Gene expression dynamics underlying diversification and expansion of an actin gene family

Edward Alan John Tunnacliffe

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London

MRC Laboratory of Molecular Cell Biology

University College London

January 4, 2018

I, Edward Alan John Tunnacliffe, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

During the evolution of gene families, functional diversification of proteins often follows gene duplication. However, some gene families expand while preserving protein sequence. Why would a cell need to maintain multiple copies of the same gene? In this thesis I have addressed this question for an actin gene family containing 17 genes encoding an identical protein in the social amoeba *Dictyostelium discoideum*. Using bioinformatics I identified several highly conserved sequence elements as potential regulatory motifs, yet found that gene expression patterns during development are broadly similar across the gene family. Turning to live cell imaging I showed that family members display different transcription dynamics, with strong 'bursty' behaviours contrasted by more steady, continuous transcriptional activity. By switching promoters I showed that different dynamics are directly determined by endogenous promoter sequences, rather than genomic context. I have explored how cell-to-cell variability in gene expression introduced by bursty transcription propagates to resultant cytoplasmic mRNA and protein and showed that population variance of these molecules is reduced compared to nascent transcription. Finally, I generated cell lines with up to 6 genes knocked out and showed that these cells potentially display a minor defect in growth. Overall these data suggest that expanded gene families are utilised not only to generate sufficient protein for normal cell physiology, but also to enable both robustness and responsiveness to a range of stimuli regulating the expression of essential genes.

Impact Statement

Understanding mechanisms of gene regulation is central to our comprehension of the basic processes of cell biology. In this work I have developed our knowledge of several processes in the field of gene expression regulation. Firstly, I showed that multiple genes encoding the same amino acid sequence can be differentially regulated at the level of transcription. Groups of genes encoding identical protein isoforms have typically been thought to be coordinately regulated and therefore this demonstrates an increased level of complexity in these systems which can be used to the cell's advantage. This differential regulation is characterised by the high level of variability in transcription of some genes compared to others. Here I showed that this variable behaviour of a gene is controlled, almost entirely, by its promoter sequence. Previous studies have implicated contextual information of the genomic locus (such as chromatin state) in the control of this process, yet here I demonstrated that the importance of such influences are minor, at least in this biological system. This promoter-controlled variability in transcription may be important for the response of a single-celled organism to its environment, which ties in with previous studies on this subject. Finally, I demonstrated that despite this variability at the level of transcription, and its potential implications for responsiveness, this noise or heterogeneity in the system is filtered out by the time protein molecules are made. This suggests that inherent robustness is built into the regulation of highly expressed housekeeping genes, perhaps by translational control, which minimises noise in the expression of these genes. Overall, these results repre-

sent an advance in our understanding of both general mechanisms of gene expression regulation, as well as the regulation of highly expressed and apparently redundant multigene families.

Acknowledgements

Firstly I would like to thank my supervisor, Jonathan Chubb, for his constant support, guidance and understanding. His unwavering optimism and enthusiasm for the project, particularly in the face of a series of failed Southern blots (the bane of my PhD life), was both inspiring and reassuring. Most importantly, I learned 'how to do science' from him, for which I will always be grateful.

I thank my lab for constructive criticism, new ideas, and delicious lab lunches on Charlotte Street and beyond; my friends at the LMCB for the many lengthy tea breaks, heated pub discussions and competitive football matches which provided much needed respite from the daily slog; and my office mates for constantly distracting me, providing me with sweets and chocolate from around the world, and making me laugh, even when I didn't feel like it.

I want to acknowledge my 'non-science' friends for getting me out of the lab and exploring London, while pretending to be interested in the science.

Finally, I want to thank my parents and my sister for always being just a phone call or train ride away and for giving me encouragement and support when I needed it most.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Evolution by gene duplication

Gene duplication has long been recognised as an important mechanism by which to achieve complexity in evolution. Early studies showed that Bar mutations in *Drosophila melanogaster* (Dobzhansky, 1932) occurred as a result of gene duplication (Bridges, 1936), linking phenotype to the duplication of template genetic material. Further work popularised the concept, with Susumu Ohno's influential book (Ohno, 1970) postulating that duplication and subsequent relaxation of selection drives genomic innovation, and therefore evolution. The role of gene duplication in the evolutionary process was cemented further in the post-genome era with it becoming clear that a large proportion of all genomes are duplicate genes (Zhang, 2003). Indeed, many of these duplicates can be further organised into multigene families which vary in size, gene conservation and genomic organisation, and the study of which can inform our understanding of the mechanisms behind genomic evolution (Walsh & Stephan, 2008). In the following work I have explored the potential contributory factors behind the expansion of a multigene family encoding identical actin proteins in the social amoeba *Dictyostelium discoideum*, primarily in the context of differential regulation within the gene family.

1.1.1 Mechanisms of gene duplication

Duplication of DNA can occur across multiple scales, from polyploidy or whole genome duplication (WGD) as observed in many yeast and plant species (and to a lesser extent in animals) (Van de Peer et al., 2017), to segmental duplication of chromosomes (Bailey et al., 2002; Koszul et al., 2006) and related small-scale duplication (SSD) events (Reams & Roth, 2015). WGD occurs via alteration of the normal cell cycle by either skipping or truncating mitosis (Yant & Bomblies, 2015). This process takes place as part of normal development in certain somatic, often terminally differentiated, cell types in many organisms, and can be beneficial for wound healing, stress tolerance and DNA damage responses, due to changes in both cell size and/or DNA content (Yant & Bomblies, 2015). However, changes in ploidy can also be adaptive. In plants and fungi, WGD can facilitate speciation as well as adaptive evolutionary responses, such as escape from herbivory or tolerance to sudden environmental stresses (Van de Peer et al., 2017). Furthermore, WGD has been implicated in many different cancers and contributes to genomic instability, a hallmark of tumour progression (Davoli & de Lange, 2011).

SSD events are thought to proceed via several mechanisms, the simplest of which is tandem duplication (Reams & Roth, 2015). Here, the process of unequal crossing over between homologous DNA sequences on closely situated chromosomes will result in the duplication of the region between homologous sites on one chromosome, with the concomitant loss of this region on the other (Fig. 1.1) (Green, 1963). An obvious example of the repetitive sequences required for such crossover are transposable elements which constitute large proportions of eukaryotic genomes and facilitate several mechanisms of genome evolution (Kazazian, 2004). Indeed, the Bar duplication described above was later shown to have occurred due to transposon-mediated recombination (Tsubota et al., 1989). Multiplication has also been proposed to occur without the need for rearrangement be-

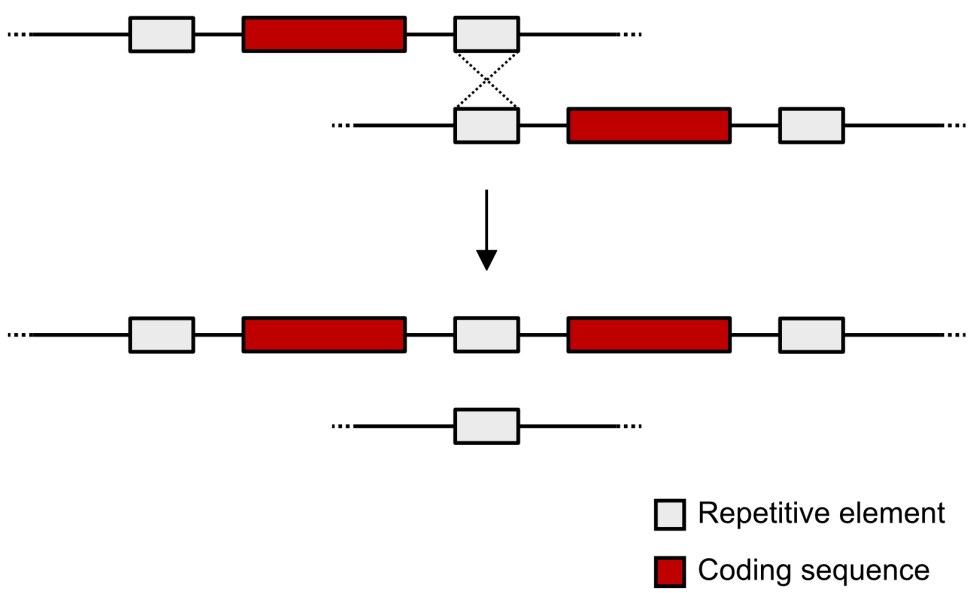

Figure 1.1: Gene duplication by unequal crossing over. Homologous recombination between similar DNA sequences, such as repetitive elements, can unequally segregate coding sequences between two chromatids or chromosomes by the mechanism shown. Multiple rounds of unequal crossing over can lead to a large tandem array of the same gene along a chromosome.

tween chromosomes. The formation of tandem inversion duplications (TID) relies upon palindromic repeats which, following a strand break and subsequent DNA repair synthesis, may be able to 'snap-back' onto the template strand, enabling triplication of the intervening sequence (Fig. 1.2) (Reams & Roth, 2015). Several other mechanistic models are described in Reams & Roth (2015).

1.1.2 Duplicate gene retention despite redundancy

Following a duplication event, gene sequences are present in (at least) two copies in the genome. Assuming that these sequences are identical and subject to the same regulatory constraints (which is not always the case, see Katju & Lynch (2006)), they will perform the same function within the cell, i.e. they are redundant. Over time, duplicate genes are typically thought to diverge in both sequence and function (Nei, 1969), yet functional redundancy, in some form, appears widespread among paralogous genes (Dean et al.,

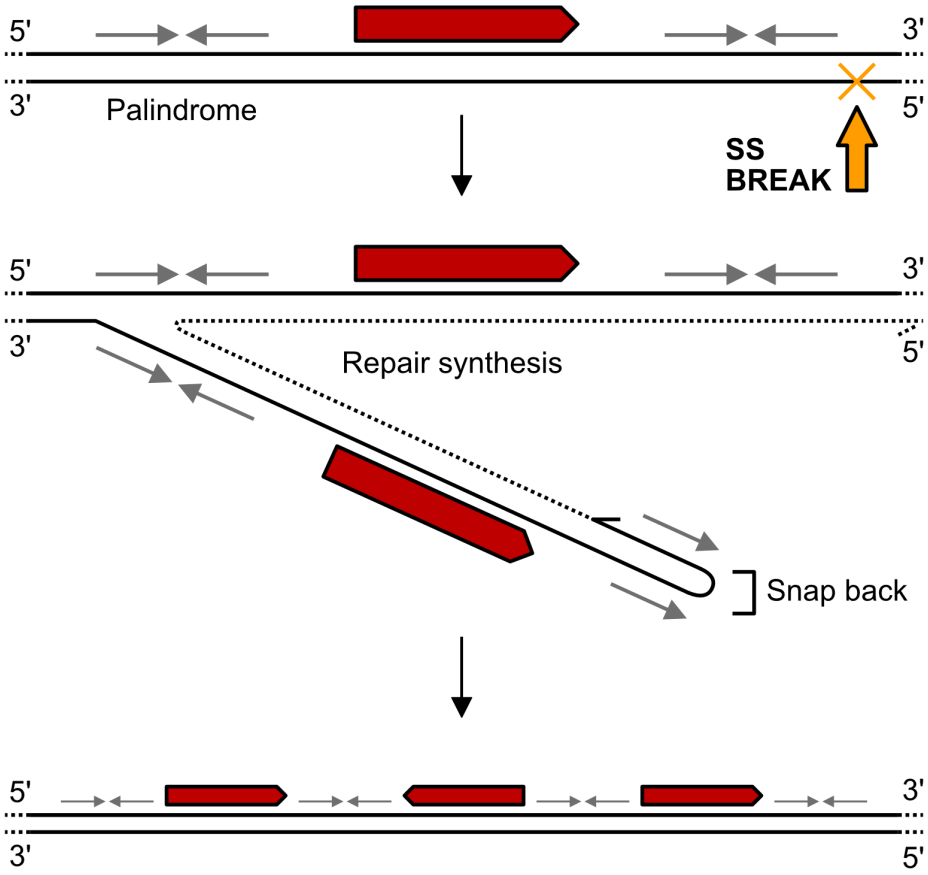

Figure 1.2: Gene duplication by tandem inversion duplication. Palindromic sequences on either side of a gene can facilitate gene duplication on a single chromosome. A single strand break at one end of the gene, close to a palindrome can cause one half of the palindrome to 'snap back' by complementary base pairing with the other half-site. This can prime DNA repair synthesis resulting in the duplication of the coding sequence. Subsequent replication of this sequence will produce one copy of the chromosome with three tandem copies of the coding sequence. Adapted from Reams & Roth (2015).

2008; DeLuna et al., 2008; Musso et al., 2008). In theory, truly redundant genes would not be protected from accumulation of deleterious mutations by purifying selection, as the very nature of redundancy precludes fitness costs upon loss of an individual paralogue (Nowak et al., 1997). In the light of this, how do redundant genes becomes fixed in a population before elimination by mutation? Numerous models have been proposed over the years to explain this phenomenon (Innan & Kondrashov, 2010); I will briefly summarise the main concepts here.

1.1.2.1 Neofunctionalisation

First proposed by Ohno (1970), this model of duplicate preservation relies on two assumptions that (1) a gene duplication event is neutral, that is, it neither positively nor negatively affects the fitness of the organism, and (2) that the duplicate is fixed in a population by genetic drift. Ohno suggested that if both of these assumptions are satisfied then constraints imposed by negative selection are relaxed on the new gene copy. This would allow the accumulation of mutations in the duplicate, and in most cases pseudogenisation occurs as a result. However, in some cases non-disruptive mutations will occur, and these can lead to the generation of novel functions for the duplicate gene. This novelty will then be subject to selection, providing a mechanism for long-term preservation of the previously redundant gene in the population. A simple example of how small changes in protein sequence can generate new selectable functions is the single mutation at residue 9 in both Bicoid and Paired homeodomain proteins which alters the specific DNA binding motif in *Drosophila* (Hanes & Brent, 1989; Treisman et al., 1989). Similarly, the mutation of just three residues can rewire the specificity of two-component signalling pathways in bacteria (Skerker et al., 2008).

1.1.2.2 Subfunctionalisation

In contrast to Ohno's theoretical model, rather than the generation of new function, subfunctionalisation allows for the maintenance of duplicate genes

by dividing the parental gene functions between the resultant duplicates. Following duplication, both paralogues are initially subject to relaxed selection enabling the segregation of parental gene functions between the two by differing accumulation of mutations. Several different models describe the fixation of genes in a population by subfunctionalisation (see DDC, EAC, permanent heterozygote, reduced expression, Innan & Kondrashov (2010); Qian et al. (2010)). These models differ only in the initial selection pressure (either neutral or positive), and the way in which the division of function is achieved. These include: loss-of-function mutations accumulating in both copies resulting in each performing separate functions of the original gene copy (duplication-degeneration-complementation, DDC); gain-of-function mutations in duplicates enabling the improvement of individual parental gene functions above that possible within a single gene (escape from adaptive conflict, EAC); fixation of two advantageous alleles of the same gene by duplication and recombination (permanent heterozygote); concerted expression reduction of both duplicates such that loss of either becomes deleterious (reduced expression). In each case, upon fixation selection acts on both duplicates as each is required to fulfil the parental gene function (Innan & Kondrashov, 2010). A simple, yet elegant example of subfunctionalisation is that of the *Hoxa1* and *Hoxb1* genes. These genes are separately controlled by two different regulatory elements which enable distinct developmental expression of these paralogues. The combination of these elements at a single locus can recapitulate the function of both genes (Tvrdik & Capecchi, 2006).

1.1.2.3 Positive dosage

Most of the above models assume only neutral selection on the duplicate gene copy. It is proposed that in some cases, duplication of a gene can be immediately beneficial for the organism and therefore will be positively selected for within a population (Innan & Kondrashov, 2010). In the case of positive dosage this refers to genes for which the increased product gen-

eration resulting from having extra copies of the gene confers an advantage. Positive selection on duplication has been demonstrated in numerous cases of adaptive responses to challenging environments (Kondrashov, 2012), with a classic example being the adaptive duplication of yeast hexose transporters in glucose-limited conditions (Brown et al., 1998). This model is also thought to be applicable to highly expressed gene families such as ribosomal genes and histones (Sugino & Innan, 2006). However, whether this mechanism applies more generally beyond these relatively specific examples is unclear (Qian & Zhang, 2008; Kondrashov, 2012).

1.1.2.4 Adaptive radiation

Another model which relies upon positive selection for newly formed duplicates is the adaptive radiation model proposed by Francino (2005). In this case, novel protein function is achieved by improving upon existing functions which are 'pre-adapted' to a particular situation. During the exploration of a new environmental niche a novel, beneficial molecular function may initially be fulfilled by suboptimal protein function—a good example is a receptor binding a new ligand with relatively low affinity—and for which a duplication could increase the level of protein and enable "biologically relevant levels of protein functionality" (Francino, 2005). Repeated gene amplification followed by diversification within a population would enable refinement of this new function, after which the most effective derivative becomes fixed and others become superfluous and subsequently pseudogenised. The well-studied example of olfactory receptor (OR) gene families is presented as a case study for the model: here, around 400 different OR genes are present in humans, of which 50% are pseudogenes (Francino, 2005; Niimura & Nei, 2006). Rapidly evolving gene families involved in predator-prey interactions, such as conotoxins (venomous peptides of *Conus* species) (Chang & Duda, 2012), are also potentially explained by this model.

1.1.2.5 Dosage balance

The concept of dosage balance refers to genes which are sensitive to changes in dosage (or levels of protein expression), and therefore gene duplication (Innan & Kondrashov, 2010). This phenomenon is widely manifested by a poor tolerance of aneuploidy compared with polyploidy (Birchler & Veitia, 2012). Dosage-sensitive proteins are often those found in protein complexes, with sensitivity derived from a requirement for specific stoichiometries within such multi-protein structures (Papp et al., 2003). As a result, duplication mechanism impacts whether paralogues are selected for under this model. The balance hypothesis predicts that SSD paralogues in complexes are less likely to be retained, while WGD paralogues—in which all complex members have been duplicated—are more likely to be retained after the event. This was demonstrated in yeast (Papp et al., 2003) and other organisms (Birchler & Veitia, 2012). Thus, the different selection pressures on duplicates generated by different mechanisms likely explains the associated functional separation of these genes (Hakes et al., 2007).

1.1.2.6 Backup buffering

Although not formalised into a testable model, a number of studies suggest that duplicate genes are able to buffer the expression of paralogous partners in response to loss or stochastic perturbation (Kafri et al., 2005, 2006; DeLuna et al., 2010). This concept is similar to, but not the same as, the buffering hypothesis that duplicates exist simply to shield against deleterious mutations (Innan & Kondrashov, 2010). Backup buffering capacity is thought to decrease with increasing sequence divergence of paralogues and therefore this may be simply be a consequence of gene duplication, unless specifically selected for in evolution (Li et al., 2010b). Furthermore, this type of genomic robustness has been linked to survival in variable environments, as those organisms in highly predictable environments do not exhibit such redundancy (Mendonça et al., 2011).

1.1.2.7 Summary

In summary, there are numerous theoretical models to explain the fixation and preservation of duplicate genes in a population, the most prominent examples of which I have outlined here. While the proponents of some of these claim that they are general models capable of explaining a large proportion of duplication fixation events, others such as the positive dosage hypothesis appear to be confined to more specific examples. However, these models are not necessarily mutually exclusive and it may be that a number of factors influence the preservation of gene duplicates, depending on the evolutionary constraints at the time of duplication. It is also worth noting the importance of considering not just the gene which is duplicated and its function, but also the fact that for long-term preservation to occur the gene must become fixed throughout the population, which can occur either by random genetic drift or positive selection.

1.2 Multigene families and their evolution

A natural consequence of generating complexity by gene duplication during evolution is the formation of multigene families. These families are clusters of genes which are often similar in both sequence and function and are predicted to have diverged, following gene duplication, from a single common ancestor (Walsh & Stephan, 2008). The size of gene families within a genome can be highly variable, from a single, unique member to hundreds of genes. The distribution of family size within a genome typically follows a power-law distribution, with relatively few families dominating the total number of genes (Huynen & van Nimwegen, 1998). Furthermore, the ratio of large to small families increases with genome size (Huynen & van Nimwegen, 1998). Simulations have shown that such power-law distributions arise even when considering only duplication and pseudogenisation, assuming these processes occur dependently within families (Huynen & van Nimwegen, 1998; Hughes & Liberles, 2008). Gene families are thought to evolve by several mechanisms (Fig. 1.3), the relative importance of which

has prompted significant controversy in the field. Here I will discuss the major mechanisms of gene family evolution, and in the process describe some of the most well-studied examples of multigene families.

1.2.1 Divergent evolution

Perhaps the simplest model of gene family evolution is divergent evolution. By this mechanism, genes within a family will slowly diverge over time with the acquisition of new functions for individual members (Fig. 1.3). The classical example of a gene family which has evolved by this mechanism is the globin gene family in vertebrates (Ingram, 1961). Here, an ancestral gene is thought to have been present before the vertebrate-invertebrate divergence, with subsequent duplications and diversification in vertebrates leading to the separation of neuroglobin, followed by myoglobin from the lineage of genes which constitute haemoglobin (Hardison, 2012). Further duplications allowed the divergence of β-globins from α-globins and subsequent specialisation, perhaps by neofunctionalisation, of these genes facilitated the evolution of separate foetal and adult haemoglobin molecules (Ingram, 1961; Hardison, 2012), enabling efficient oxygen transfer between mother and child.

However, in the 1970s studies of rDNA in *Xenopus* spp. made it clear that simple divergence could not explain the evolution of all multigene families. In particular, the observation that rDNA genes were more similar within species than between species led to the idea that individual gene families could evolve cooperatively (Brown et al., 1972) (see below). Despite this, the principles of divergent evolution as a mechanism for generation of novelty still form a key part of multigene family evolution.

1.2.2 Concerted evolution

As described above, the study of rDNA genes led to the concept of concerted evolution, where gene families evolve ‘in concert’ such that the nucleotide sequences are homogenised within a particular gene family. rDNA

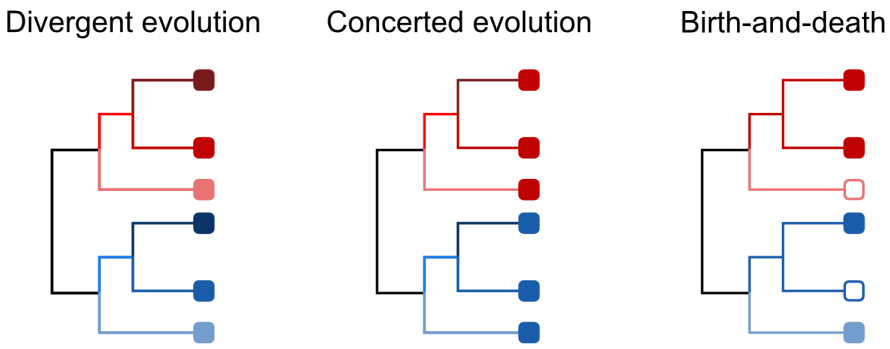

Figure 1.3: Models of gene family evolution. Schematics show simplified phylogenetic trees for each model. Squares are individual gene duplicates, with shades of colour representing sequence similarity. Open squares are pseudogenes. Divergent evolution predicts gradual divergence of function from a single ancestral gene. Concerted evolution states that duplicates evolve 'in concert' by homogenisation of sequences. The birth-and-death model allows both divergence within a family, as well as maintenance of sequence by purifying selection while predicting a large number of pseudogenes within the family. Adapted from Nei & Rooney (2005); Futuyma (2005).

genes encode the four rRNAs which, along with around 80 ribosomal proteins (Kenmochi et al., 1998), make up the 80S ribosome in eukaryotes (Ben-Shem et al., 2011). While the ribosomal proteins are generally encoded by single copy genes, rDNA genes are often present in several hundred copies within the genome (Eickbush & Eickbush, 2007), and up to tens of thousands in some species, with the variability predominantly explained by genome size (Prokopowich et al., 2003). Three of the four rRNAs, along with transcribed spacer regions (TS), are encoded by a single gene or transcription unit, which is subsequently processed into individual RNAs (Wellauer & Dawid, 1977; Eickbush & Eickbush, 2007). Typically, these genes are arranged in tandem arrays of hundreds of copies with intergenic spacer (IGS) regions between (Miller & Beatty, 1969). Brown et al. (1972) observed that the coding and IGS sequences of rDNA genes were almost identical between two closely related *Xenopus* species. However, inter-species comparisons of TS sequences showed significant divergence, despite high intra-species similarity (Brown et al., 1972). This indi-

cated that while TS (but not coding) sequences appeared free to diverge between species, within a species these sequences were somehow constrained, which couldn't be explained by simple divergence alone. As such, the concept of concerted evolution was born.

What is the mechanism behind concerted evolution? Initially, unequal crossover (Fig. 1.1) was proposed as a mechanism by which homogeneity of rDNA sequence could be maintained, given that members of the same species showed variable numbers of rDNA genes, and that individuals retained many more copies of the genes than required for survival (Eickbush & Eickbush, 2007). Computer simulations and experimental work in yeast supported this view (Ohta, 1976; Petes, 1980; Szostak & Wu, 1980) but suggested that unequal crossover was not the only mechanism driving concerted evolution. A complementary mechanism, gene conversion, was first proposed by Szostak et al. (1983). This involves the non-reciprocal transfer of DNA from a 'donor' sequence to a highly similar 'acceptor' sequence (Chen et al., 2007). The process is thought to occur between sequences with $>$95% homology after a double strand break event (see figure 1.4 for a simplified mechanism), with conversion tracts about 200 bp - 1 kb in humans (longer in yeast) (Chen et al., 2007). The combination of this mechanism with unequal crossover, provided a satisfactory theoretical outline for concerted evolution of rDNA genes.

The popularity of this model led to the assumption that many multigene families had evolved by concerted evolution. Indeed, researchers used this model to explain the evolution of MHC, immunoglobulin, tRNA, heat shock protein, olfactory receptor and other gene families (Brown & Ish-Horowicz, 1981; Ohta, 1983; Sharon et al., 1999). Even a few globin genes have since been shown to have undergone some level of gene conversion (Hardison, 2012). In particular, another well-studied gene family, the histones, was proposed to have evolved analogously to rDNA given both sets of genes are highly expressed and often arranged in tandem (Nei & Rooney, 2005). Ini-

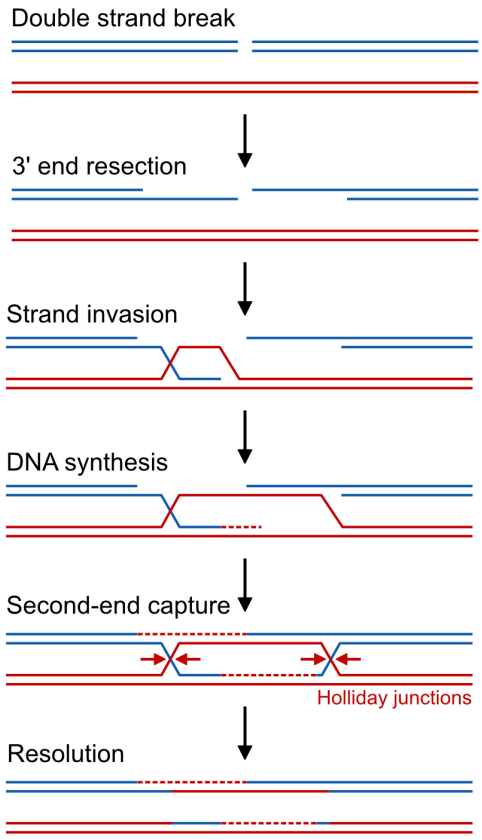

Figure 1.4: Basic mechanism of gene conversion. Following a double-strand break, 5' to 3' exonuclease cause 3' end resection of the free ends of the DNA. The resulting ssDNA tails are able to interact with nearby homologous dsDNA by strand invasion, which acts as a primer for DNA synthesis to occur. This synthesis extends the newly formed D-loop (displaced red strand) until the other resected end is reached and captured by ligation. This ultimately results in the formation of two Holliday junctions, which are resolved into the final gene conversion product. Adapted from Chen et al. (2007).

tial studies focussed on the 'early' histone genes in sea urchins, i.e. those expressed from late oogenesis to the blastula stage of the embryo. This family has a highly regular arrangement with all five histone genes clustered into a tandem array, which itself is tandemly repeated hundreds of times at several genomic loci (Cohn & Kedes, 1979; Holt & Childs, 1984). Based on restriction mapping and limited DNA sequencing of repeat units it was determined that this gene family had evolved by concerted evolution (Holt & Childs, 1984). The acquisition of more extensive sequence data from

several clones in *Drosophila* compared divergence within and between individuals and species and concluded that the low rate of variability in individuals was also indicative of concerted evolution (Matsuo & Yamazaki, 1989). However, the limited amount of nucleotide sequence data and reliance on restriction enzyme mapping of gene families meant that the importance of purifying selection on the maintenance of histone coding sequences was underappreciated at the time. A new model of gene family evolution, known as the birth-and-death model, was proposed to explain the unusual patterns of divergence in MHC loci (Hughes & Nei, 1988, 1989a) and would later be shown to explain the evolution of most other gene families, including histones (Nei & Rooney, 2005).

1.2.3 Birth-and-death model

As described above, the prevailing view in the mid-1980s was that evolution of multigene families was primarily dictated by concerted evolution. MHC genes were seemingly no exception to this rule, despite the fact that these genes are highly polymorphic and concerted evolution had been demonstrated to be a mechanism by which to homogenise related sequences within the genome (Eickbush & Eickbush, 2007). Gene conversion was thought to introduce polymorphisms by shuffling smaller portions of coding sequence between genomic loci within individuals (Weiss et al., 1983; Ohta, 1983). However, these arguments were based on comparisons of small numbers of sequences considering amino acid and total DNA sequence divergence without allowing for the role of selection in the process. The work of Hughes and Nei (Hughes & Nei, 1988, 1989b,a) demonstrated selection for adaptive mutations in these genes, which wouldn't be expected by concerted evolution, by comparing the ratio of non-synonymous to synonymous mutations. Furthermore, they showed that MHC genes are derived from an ancestral gene by a series of nucleotide substitutions, and that further substitutions lead to formation of pseudogenes (Hughes & Nei, 1989a).

These observations formed the basis of the birth-and-death model of

gene family evolution (Nei & Rooney, 2005). As the name suggests, gene families are proposed to be in constant flux with their size increasing and decreasing by 'birth' and 'death' of genes over time. Genes are 'born' by duplication and are free to accumulate mutations over time. Some of these mutations will be deleterious and lead to gene 'death' while other genes are maintained in the genome for long periods of time by selection. This model provided a simple explanation for the diversification of MHC genes and was subsequently applied to many other diverse gene families including immunoglobulins (Nei et al., 1997), T-cell receptors (Su et al., 1999) and chemosensory receptors (Niimura & Nei, 2003; Vieira et al., 2007) among others (Nei & Rooney, 2005). The model can also accommodate strongly uniform gene families, with purifying selection enabling the maintenance of particular coding sequences. This has been demonstrated for ubiquitins (Nei et al., 2000), actins (Zhu et al., 2013) and, contrary to previous studies, histones (Piontkivska et al., 2002; Rooney et al., 2002; Eirín-López et al., 2004). Given the weight of evidence in its favour the birth-and-death model is now thought to explain the evolution of most, but not all, gene families (Nei & Rooney, 2005; Eirín-López et al., 2012).

1.2.4 Summary

The models outlined above attempt to explain how gene families are shaped over time. Despite the birth-and-death model having emerged as perhaps the dominant force in gene family evolution, it is important to recognise that the models are not necessarily mutually exclusive. Indeed, Martinsohn et al. (1999) conclude that gene conversion has occurred between MHC genes, but that it likely occurs without adaptive significance. Similarly, even rDNAs, the classic example of concerted evolution, have been shown to follow a birth-and-death model in filamentous fungi (Rooney & Ward, 2005). The careful analyses clearly required for these kinds of studies have been made easier in the post-genomic era and will benefit further from the comparison of numerous closely-related species (Vieira et al., 2007; Zhu et al., 2013).

1.3 Actin gene families in the kingdoms of life

The subject of this thesis is the actin gene family of the social amoeba *Dictyostelium discoideum*. Before introducing the organism and gene family specifically, I will briefly review current knowledge of actin gene families in other organisms from diverse backgrounds, with regards to their size, evolution and regulatory differences.

Actin is a highly conserved protein found in all eukaryotic cell types (with the exception of nematode sperm) (Gunning et al., 2015; Roberts & Stewart, 1997). The protein was initially identified from its role in muscle contraction and was subsequently found to be required for a wide range of functions in almost all cell types including, but not limited to, cell shape changes, division, motility, junction formation, vesicle trafficking and perhaps transcriptional regulation (Perrin & Ervasti, 2010). This diversity of function has emerged despite strong conservation of sequence both within and between species (Gunning et al., 2015). Gunning et al. (2015) proposed that different kingdoms have evolved such functional diversity via several different mechanisms, the understanding of which requires consideration of both actin and its many binding partners.

1.3.1 Animals

Actin gene families of most animals are relatively small, often containing fewer than 10 members with each gene encoding a unique protein isoform (Gunning et al., 2015). Of these, there are generally two forms of cytoplasmic actin, β and γ, which are expressed in most different cell types. The remaining actin genes encode proteins expressed in various different muscle types —cardiac (α_{cardiac}), skeletal (α_{skeletal}), vascular smooth (α_{smooth}), and enteric smooth (γ_{smooth}). Amino acid sequence divergence within individual animal families is minimal, with all actin isoforms at least 93% identical in mammals and birds, while β- and γ_{cyto}-actin differ by only four amino acids (Perrin & Ervasti, 2010). Furthermore, actin proteins are highly conserved between species, with cytoplasmic isoforms 100% identical in mammals

and birds, for example. This sequence similarity suggests strong purifying selection acting on protein sequences. Indeed, a phylogenetic analysis of actin genes in primates showed that this, coupled with the presence of many actin pseudogenes in animal genomes, suggests that actin gene families are subject to the birth-and-death model of evolution (Zhu et al., 2013). Actin-related (Arps) and actin-like proteins show more limited homology to conventional actins, but orthologous Arps are themselves highly conserved across species (Schafer & Schroer, 1999).

Despite the high sequence similarity of actin genes, both unique and overlapping functions have been ascribed to different actin isoforms in animals (Perrin & Ervasti, 2010). The majority of information regarding these functions comes from knockout mouse models. Predictably, muscle-specific actins are expressed in muscle cell types and knockout mice exhibit varying phenotypic severity. The majority of mice lacking α_{cardiac}- or α_{skeletal}-actin die in early development (Kumar et al., 1997; Crawford et al., 2002) while α_{smooth} knockout mice are viable but show defects in vascular contractility and blood flow (Schildmeyer et al., 2000). Of the cytoplasmic actins, which are ubiquitously expressed, β-actin null mice are embryonic lethal (Bunnell et al., 2011), while γ_{cyto}-actin knockouts are viable but show defects in growth and survival (Belyantseva et al., 2009; Bunnell & Ervasti, 2010). In all cases, disruption of particular actin genes induces upregulation of others in the gene family. The fact that defective phenotypes are still evident in these mice suggests that these effects may be due to the loss of specific actin isoforms which perform unique functions, and not reduced protein levels. In keeping with this, while transgenic expression of α_{cardiac}-actin can rescue lethality of an α_{skeletal}-actin knockout (Nowak et al., 2009), presumably due to their 99% sequence identity and overlapping expression patterns, overexpression of γ_{cyto}-actin in the α_{skeletal}-actin knockout was unable to rescue the phenotype (Jaeger et al., 2009).

Therefore, functional diversity has been achieved, to some extent, by

expansion of the actin multigene family in animals. The molecular mechanisms of this diversity are likely twofold. Clearly, muscle actins are tissue-specific, but cytoplasmic actins can also be differentially localised within the cell with the RNA-binding 'zip-code' motif in the 3' UTR of β-actin a contributory factor (Kislauskis, 1994; Dugina et al., 2009). Furthermore, certain actin-binding proteins (ABP) have been shown to discriminate between muscle and cytoplasmic isoforms of actin (Perrin & Ervasti, 2010). These attributes likely enable diversification of action for the individual actin isoforms, which collectively contribute to the wide-ranging functions that actin performs within animals.

1.3.2 Plants

The number of actin genes in plants is considerably more variable than in animals. Algae typically have two actin genes, one conserved and one variable in sequence (Wu et al., 2009). Multicellular plants have additional copies, with several species containing around 10 genes, while soybean (17) and maize (21) appear to have more (Šlajcherová et al., 2012; Gunning et al., 2015). An earlier study in petunia identified a large gene family of several hundred members by screening a phage genomic DNA library by hybridisation (Baird & Meagher, 1987), although it is unclear how many of these might be pseudogenes. Overall, most complex plants surveyed to date appear to have larger actin gene families than animals.

Could this have enabled a greater diversification of function too? The most detailed studies of actin genes in plants were done on *Arabidopsis thaliana*. The family of 10 genes contains two pseudogenes, and can be subdivided further into vegetative and reproductive actin genes (Šlajcherová et al., 2012). The reproductive genes may have evolved from the more primitive vegetative genes (An et al., 1999). Of the eight active isoforms in *A. thaliana* there are three pairs of closely-related paralogues, of which two genes are identical, and the other two paralogous pairs differ by only one amino acid each (Šlajcherová et al., 2012). However, these paralogous

pairs show nucleotide sequences which have diverged by 9-12% which suggests that purifying selection has been in action over considerable evolutionary distances (Šlajcherová et al., 2012). Furthermore, plant actins show higher divergence within the gene family compared to animals (Gunning et al., 2015) and divergent amino acids are more likely to be at the surface of the molecule, potentially signifying modulation of binding partner interaction (McDowell et al., 1996).

The regulation of *A. thaliana* actin gene expression is in line with the relationships outlined by the phylogenetic tree—vegetative and reproductive actins are expressed in their respective tissue types, while paralogous pairs are similarly expressed (Šlajcherová et al., 2012). For the vegetative genes, the paralogues ACT2 and ACT8 were shown to be able to compensate for each other's loss, but the more distantly-related ACT7 had a unique function and was unable to fully rescue a specific root hair phenotype (Kandasamy et al., 2009). Actin-binding proteins also appear to have an important role in the functional diversification of plant actin genes, as profilins are similarly specialised as either vegetative or reproductive proteins (Huang et al., 1996). It was later shown that defects associated with ectopic expression of a reproductive actin in vegetative tissues could be mitigated only if reproductive profilins were co-expressed in these cells (Kandasamy et al., 2007). In summary, actin gene families in plants are larger and more diverse than in animals, with co-evolution of ABPs seemingly important for the functional divergence of actin isoforms.

1.3.3 Fungi

In comparison to animals and plants, fungi have a relatively simple actin cytoskeleton. Yeasts and other fungi have a single actin gene, as well as single or very few copies of binding proteins such as ADF, profilin, formin and tropomyosin (Gunning et al., 2015). Despite this relative lack of complexity, fission yeasts are still able to generate diversity of actin function by organising the protein into three distinct structures—actin patches, rings

and cables—each of which have a unique set of interacting binding proteins (Kovar et al., 2011).

1.3.4 Bacteria

While actin proteins are specific to eukaryotes, bacterial cells also require protein polymers to perform similar force-generating functions. However, actin-like functions in bacteria are divided between three separate proteins—ParM, MreB and FtsA (Gunning et al., 2015). Each of these have separate functions with ParM segregating large DNA plasmids (Jensen & Gerdes, 1997), MreB being important for cell shape regulation and cell wall synthesis (Doi et al., 1988) while FtsA combines with the tubulin homologue FtsZ to form the Z-ring (Addinall & Lutkenhaus, 1996). Similar to eukaryotic systems, the level of phylogenetic divergence of these actin-like molecules appears to be linked to the array of interacting partners associated with each protein. Indeed, ParMs are the most divergent proteins across bacterial species (Gunning et al., 2015), which could in part be due to the fact these molecules only bind themselves and a filament end-binding protein (ParZ) (Salje et al., 2010). Therefore there are likely to be fewer selective constraints on protein sequence and structure than MreB and FtsA which have many interacting partners and are also considerably more similar between species than ParM (Gunning et al., 2015). The evolution of this actin-like network represents a different approach to the ‘one-size-fits-all’ strategy employed by actin in eukaryotes, where bacterial cells have evolved distinct proteins for separate functions within the cell. The recent discovery of actin-like molecules in archaea (Lindås et al., 2017) may reveal further strategies by which to coordinate cytoskeletal function in different organisms.

1.3.5 Summary

Organisms from all kingdoms utilise actin-like protein polymers to generate force and provide structural support for a diverse range of cellular processes. The mechanisms by which this diversity of function is achieved

varies with evolutionary divergence. Bacterial systems use several divergent actin-like proteins to achieve unique functions within the cell. This system surely has some advantages compared to use of a single class of actin protein in eukaryotes, such as the relative freedom to evolve individual protein sequence without affecting the function of the others. However, by maintaining a single highly conserved gene family, eukaryotes ensure that only one pool of actin-like molecules is required to carry out numerous cellular functions. Of course, this alternative approach requires increased regulatory complexity and eukaryotes seem to have taken multiple routes to achieve this. Gunning et al. (2015) note that while plants have expanded the number of actin genes and actin binding proteins like ADF and profilin in order to generate diversity, fungi and animals appear to instead utilise tropomyosins to broaden the functional scope of each actin isoform. Indeed, plants (and protists) appear to completely lack tropomyosin homologues (Gunning et al., 2015). This highlights the importance of studying gene family organisation in order to understand mechanisms by which diverse organisms generate complex systems.

1.4 Introduction to *Dictyostelium* biology

First discovered in North Carolina, USA by Raper (1935), the social amoeba *Dictyostelium discoideum* has proved to be an incredibly useful organism for studying a diverse range of cellular processes in eukaryotes. In this section I will describe general features of *Dictyostelium* biology and discuss the merits of using this organism in experimental research, with particular regard to understanding gene expression regulation.

1.4.1 The life cycle of *Dictyostelium discoideum*

Perhaps the most striking feature of *Dictyostelium* biology is the remarkable transition it makes from a unicellular to multicellular organism in response to starvation. When food sources are plentiful, *D. discoideum* survives as a single cell in soils and decaying vegetative matter, primarily feeding on

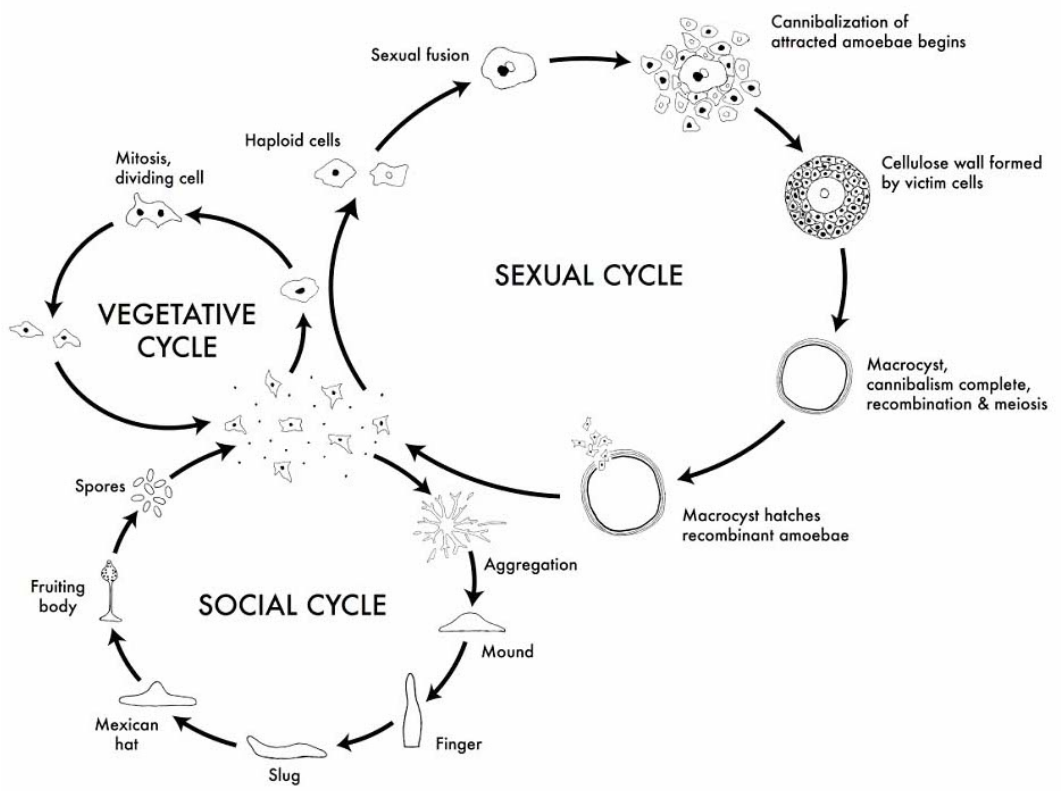

Figure 1.5: Life cycles of *Dictyostelium discoideum*. Image reproduced from dictybase.org with permission (Creative Commons).

bacterial prey (Raper, 1937). The amoebae hunt bacteria by sensing and responding via chemotaxis to secreted folate molecules (Pan et al., 1972), and engulf them by phagocytosis (Raper, 1937). Upon depletion of appropriate food sources, *Dictyostelium* cells begin to secrete cyclic adenosine monophosphate (cAMP), embarking upon one of two separate life cycles, detailed below (see Fig. 1.5).

The sexual cycle of *Dictyostelium discoideum* is favoured in dark and very humid or submerged conditions (Bloomfield, 2013), where individual *Dictyostelium* amoebae, which are haploid, fuse with members of an opposite mating type (Bloomfield et al., 2010). The diploid progeny are known as giant cells and secrete cAMP to attract other amoebae via chemotaxis (O'Day, 1979; Bloomfield, 2013). Giant cells engulf individual amoebae in order to accumulate nutritional resources for macrocyst formation (Filosa & Dengler, 1972) (Fig. 1.5). This structure is a type of spore which is highly

resistant to environmental stress and enables survival of the organism until food resources are replenished, after which germination occurs and cells reenter the vegetative life cycle (Blaskovics & Raper, 1957). Macrocyst formation has been reported across all of the main dictyostelid phylogenetic groups (Bloomfield, 2013). However, beyond description of the mechanics of the process and the environmental conditions required, as well as identification of the mating locus, the sexual cycle in *D. discoideum* is relatively poorly characterised compared to the asexual life cycle described below (Bloomfield, 2013).

The social or developmental life cycle has been studied extensively since Raper's discovery of *D. discoideum* in 1935. Again, the initiation and progression of this programme is controlled by cAMP. Upon depletion of food supplies, cells begin secreting cAMP which is sensed by neighbouring amoebae (Saran et al., 2002). These in turn begin to secrete cAMP which is further propagated to nearby cells, while chemotaxing towards the initial source of the signal (Bonner, 1947) (see aggregation, Fig. 1.5). This process results in a signalling relay between cells which, over time, leads to the stochastic emergence of aggregation centres by initiating waves of cellular movement within the population (Tomchik & Devreotes, 1981). During this process cells become polarised (Swanson & Taylor, 1982) by structural reorganisation of the actin cytoskeleton and other signalling components at the leading edge of the cell (Kay et al., 2008). Cells move chemotactically in a head-to-tail arrangement towards an aggregation centre (Bonner, 1947), typically by formation of pseudopods which are highly plastic actin-based extensions (van Haastert, 2010). Cells will 'stream' towards such a centre over distances of several millimetres, with tens or even hundreds of thousands of cells in a single aggregation territory.

This cAMP-mediated aggregation phase of development (Fig. 1.5) takes place during the first 6-8 hours after starvation has been initiated (Chisholm & Firtel, 2004). Streams of cells merge at the centre of the ag-

gregation territory to form a mound (Fig. 1.5). Further compaction of the mound precedes extension from the substrate into a 'finger'-like structure at about 12-14 hours of development. From here, the organism can develop into a pseudoplasmodium or 'slug' structure which is motile, polarised and both thermo- and phototactically responsive (Fig. 1.5) (Raper, 1940). Alternatively, if conditions are suitable, the organism can proceed immediately to the culmination phase of the life cycle which takes place from around 18-24 hours after starvation (Chisholm & Firtel, 2004). Here, the organism anchors itself and begins to extend a sorus of spores away from the substrate by a stalk-like projection, resulting in a fruiting body (Fig. 1.5) (Raper, 1935). Upon dispersal of these spores to new environments with plentiful food supplies, germination occurs and cells reenter the vegetative cycle (Raper, 1935). During this process, from the mound through slug stages, cell type differentiation occurs in order to generate the spore and stalk cells from previously undifferentiated amoebae. Differentiation begins in the mound stage, where cells begin to express pre-spore and pre-stalk markers in a 'salt-and-pepper' fashion, before subsequent 'sorting out' of cell types in the slug (Kay & Thompson, 2009). Stalk cells comprise around 15-20% of the total number in the fruiting body (Bonner & Slifkin, 1949) and die during the formation of the final structure.

1.4.2 *Dictyostelium* as an experimental organism

The utility of *Dictyostelium* species in biomedical research has grown over the last decades to the point that experimental understanding of this organism has augmented many different fields in biology. Given the processes involved in the life cycles described above, the study of *Dictyostelium* has contributed to our understanding of phagocytosis, macropinocytosis, differentiation, cell motility, chemotaxis, cell adhesion and social evolution (Annesley & Fisher, 2009; Williams, 2010). However, because of its numerous similarities to mammalian cell types, *Dictyostelium* has also been used to study host-pathogen interactions and neurological diseases (Annesley &

Fisher, 2009) as well as more fundamental processes such as gene expression (Chubb et al., 2013), DNA repair, autophagy and regulation of the actin cytoskeleton (Eichinger & Rivero, 2013).

Initial work by pioneers in the field such as Kenneth Raper and John Bonner in the decades following discovery of *Dictyostelium discoideum* required bacterial food sources for the propagation of cell cultures (Raper, 1937). An important advancement was made by Sussman & Sussman (1967) who developed a strain capable of axenic growth in liquid medium (while still able to phagocytose bacteria). This strain (Ax1) and those derived from it (Ax2, Ax3, Ax4) survive by increased fluid uptake by macropinocytosis, and much later it was shown that this was due to a mutation in the RasGAP NF1 (Bloomfield et al., 2015). Subsequent development of transformation techniques enabled the selection of genetically modified organisms by homologous recombination (Nellen et al., 1984) which led to specific targeting of individual genes (De Lozanne & Spudich, 1987; Manstein et al., 1989), as well as random mutagenesis using restriction enzyme mediated integration (REMI) (Kuspa & Loomis, 1992). The completion of the genome project (Eichinger et al., 2005) and the curation of a dedicated database for sequencing projects, dictyBase (Basu et al., 2013) has provided researchers with a wealth of bioinformatic data with which to explore the biology of the organism.

As a model for studying general features of eukaryotic biology, the advantages of *Dictyostelium* are numerous. The small (34 Mb) genome of *D. discoideum* (Eichinger et al., 2005) is readily amenable to genetic modification, as described. Rapid screening of mutants is facilitated by the haploid genome and the relative simplicity of generating large cell numbers in liquid culture. Compared to metazoans, dictyostelids are relatively simple organisms, having diverged from the animal lineage before fungi (Eichinger et al., 2005). Yeasts such as *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* have been extensively employed as simple eukaryotic

model organisms to much success (Botstein & Fink, 2011) but some important features of metazoan cells are absent from these cells such as motility and an unobstructed plasma membrane. Furthermore, many proteins absent in yeast and initially thought to be unique to metazoa have subsequently been identified in *D. discoideum*, including 24 classes of protein kinase (Annesley & Fisher, 2009), demonstrating more extensive conservation of some of the fundamental features of metazoan cell biology in this organism. As such, *D. discoideum* represents an important model organism of intermediate complexity for the study of basic cell biology as well as for multicellular processes in development.

1.4.3 Studying gene expression regulation in *Dictyostelium*

As with most well-studied organisms, numerous techniques for studying gene expression are available in *D. discoideum*. Traditional techniques such as *in situ* hybridisation and northern and western blotting are simple to carry out, while generation of fluorescent fusion proteins has provided a more quantitative means by which to study gene expression (Eichinger & Rivero, 2006). Of particular importance are genome-wide methods which have enabled profiling of global changes in gene expression associated with various processes, the most notable of which being the progression of the developmental life cycle. Both microarrays (Van Driessche et al., 2002) and more recently RNA sequencing (RNAseq) (Rosengarten et al., 2015) have been used to monitor the transcriptome during *D. discoideum* development. These methods have been used to compare developmental expression between *Dictyostelium* species (Parikh et al., 2010) as well as to assess changes in non-coding RNA abundance (Rosengarten et al., 2017). The data are accessible via a user-friendly web application, dictyExpress (Rot et al., 2009).

While genome-wide measurements are extremely useful, both microarrays and RNA sequencing have limitations, namely cross-reactivity of probes and issues with mapping of sequence reads, respectively. Further-

more, these methods use homogenised populations of cells, thus losing potentially valuable information from analysis of single cells. Therefore specific targeting of individual genes in individual cells is still an important approach to understand gene expression regulation. Highly quantitative methods enabling the counting of individual mRNA transcripts in single cells such as single molecule RNA fluorescence *in situ* hybridisation (RNA FISH) have been developed (Femino et al., 1998), and these methods can easily be applied to *Dictyostelium* (Corrigan et al., 2016). Single cell RNA sequencing (scRNAseq) has also been employed to measure global changes in populations of single cells during development (Antolović et al., 2017). However, all of the above methods lack the ability to visualise dynamic changes in gene expression over short periods of time. Studies using live imaging techniques in *Dictyostelium* showed that transcription of a gene is not continuous, but rather that mRNA production occurs in short bursts of activity, the relative frequency and size of which determine the extent of gene expression (Chubb et al., 2006). Since then, this technique, which utilises the MS2 system (Bertrand et al., 1998), has been used to demonstrate differential responses of individual *Dictyostelium* cells to an inducible developmental stimulus (Stevense et al., 2010), to explore the regulation of this bursting phenomenon during development (Muramoto et al., 2012; Corrigan & Chubb, 2014), and to develop a detailed quantitative modelling approach to understanding transcriptional mechanism (Corrigan et al., 2016). The wide variety of techniques available, both highly quantitative and capable of capturing single cell variation, combined with the aforementioned genetic tractability, greatly simplifies the study of gene expression regulation in *Dictyostelium*.

1.5 The actin gene family in *Dictyostelium discoideum*

Having previously discussed actin gene families in other kingdoms (section 1.3) I will now discuss actin gene families of protist species including *D. discoideum* followed by a more detailed review of the current knowledge of the gene family in this organism.

While 'protist' is essentially an umbrella term which includes a hugely diverse group of organisms, a comparison of actin gene families within this group highlights some similar strategies for genomic organisation of the cytoskeleton among more closely related species. Again, the data collated by Gunning et al. (2015) provides useful reference here. Two distantly related protists (Pánek et al., 2016) *Phaeodactylum tricornutum*, a diatom alga, and *Trypanosoma brucei*, a flagellate parasite of vertebrate animals, encode very few actin genes (three and two respectively), each as single copy genes. Similar to yeast species, they also encode either single or few copies of actin cytoskeletal regulators such as ADF, profilin and formin (Gunning et al., 2015). However, unlike yeast they lack tropomyosin proteins, suggesting that overall the actin cytoskeleton in these organisms requires little complexity in terms of its function. Conversely, other protists such as *Entamoeba histolytica* and *Dictyostelium discoideum*, while also lacking tropomyosin, have many more actin-binding regulatory proteins as well as several copies of actin itself (Gunning et al., 2015). Perhaps more curious is the fact that both these species encode a single protein isoform in multiple gene copies. Specifically, *E. histolytica* has one unique amino acid sequence for actin encoded by seven different genes, while *D. discoideum* have eight (potentially more) unique amino acid sequences, of which one is encoded 17 times within the genome (Gunning et al., 2015). Analysis of actin genes in *Arcella* spp. similarly identified a large number of genes encoding identical amino acid sequences, albeit by PCR (Lahr et al., 2011). The fact that these three organisms are members of the Amoebozoa (Pánek et al., 2016) suggests

that this might be a general feature of amoeboid genomes and marks a departure from the otherwise similar genomic organisation of cytoskeletal elements in plants (see section 1.3.2).

The phenomenon of encoding a single protein isoform in multiple genes, although rare, is not entirely unique. For example, the aforementioned rDNA genes are homogenised by concerted evolution (Eickbush & Eickbush, 2007), while there are more than 10 genes for histone H3 and H4 proteins in mice and humans which encode very few (three and one, respectively) different isoforms (Marzluff et al., 2002). Similarly, ubiquitin genes in many species, including *Dictyostelium*, are under strong purifying selection and contain the same amino acid sequence (Nei et al., 2000). To my knowledge, no other examples of this type of gene family organisation have been reported in *D. discoideum*.

1.5.1 Actin gene family architecture in *Dictyostelium*

Before examining historical studies of the regulation of the gene family, a more detailed look at the genomic architecture of the family as outlined by Joseph et al. (2008) is useful. In their analysis, Joseph et al. (2008) identified 16 different actin isoforms (determined as proteins with a 'complete actin domain profile') compared to eight by Gunning et al. (2015) as well as a unique actin variant, filactin (Fig. 1.6). The differences between the two studies may be due to the fact that some genes identified as actin by Joseph et al. (2008) have diverged significantly in sequence. Act33, for example, shares only 31% identity and 49% similarity in terms of amino acid sequence with the major actin isoform, and contains long polyN and polyS sequence tracts. Furthermore, those excluded by Gunning et al. (2015) have very few or even no ESTs registered at dictyBase suggesting these genes are not expressed and may be pseudogenes, in addition to the seven identified by Joseph et al. (2008). Therefore, the definition of an actin protein in Joseph et al. (2008) is perhaps less strict than that of Gunning et al. (2015).

Proteins with complete actin domain profile

#	Protein	Gene	dictyBase ID	Swiss-Prot ID
1	Act3	*act3*	DDB0220458	P07829
2	Act8*	*act8*	DDB0216213	P07830
3	Act10	*act10*	DDB0220457	Q54GX7
4	Act17	*act17*	DDB0185125	Q554S6
5	Act18	*act18*	DDB0220459	P07828
6	Act22	*act22*	DDB0220460	Q553U6
7	Act23	*act23*	DDB0220461	Q55EU6
8	Act24	*act24*	DDB0220462	Q54HF1
9	Act25	*act25*	DDB0220463	Q54HF0
10	Act26	*act26*	DDB0220464	Q55CU2
11	Act27	*act27*	DDB0229353	Q54HE9
12	Act28	*act28*	DDB0229354	Q54HE7
13	Act29	*act29*	DDB0229355	Q54L54
14	Act31	*act31*	DDB0234013	Q55DY5
15	Act32	*act32*	DDB0234014	Q55DS6
16	Act33	*act33*	DDB0234012	Q54JL1
17	Filactin	*fia*	DDB0220465	Q54PQ2
18	Arp1	*arpA*	DDB0220489	Q54I79
19	Arp2	*arpB*	DDB0185179	O96621
20	Arp3	*arpC*	DDB0219936	P42528
21	Arp4	*arpD*	DDB0233063	Q54UQ7
22	Arp5	*arpE*	DDB0234009	Q54E71
23	Arp6	*arpF*	DDB0234010	Q54KZ7
24	Arp8	*arpG*	DDB0234011	Q54JV5
25	Arp11	*arpH*	DDB0233828	Q54JY2

Identical actins of the Act8 group

1	Act1	DDB0220444
2	Act2	DDB0185124
3	Act4	DDB0220448
4	Act5	DDB0220447
5	Act6	DDB0185126
6	Act7	DDB0220445
7	Act8	DDB0216213
8	Act9	DDB0220456
9	Act11	DDB0220449
10	Act12	DDB0216214
11	Act13	DDB0220454
12	Act14	DDB0220455
13	Act15	DDB0185015
14	Act16	DDB0185127
15	Act19	DDB0220446
16	Act20	DDB0220450
17	Act21	DDB0220451

Figure 1.6: *Dictyostelium discoideum* actin and actin-related proteins. According to Joseph et al. (2008) *D. discoideum* contain 16 unique actin isoforms (yellow), one of which is encoded by 17 different genes (blue) known as the *act8* group of genes. Reproduced from Joseph et al. (2008).

Of those identified by Gunning et al. (2015) as actin isoforms, three (Act3, Act10, Act22) have very few amino acid substitutions (11, 1 and 4 respectively) compared to the Act8 group of proteins, while the others (Act17, Act18, Act24, Act25) have many more (between 46-69 with additional insertions and deletions). As in other systems, this diversity of amino acid sequence could be important for mediating different functions of the actin isoforms. The *act8* group of genes encoding the same protein isoform (Fig. 1.6) are distributed across the genome on several different chromosomes (Joseph et al., 2008). Phylogenetic analysis of the coding sequences suggested a series of five major duplications during the expansion of the subfamily, with further subsequent duplications producing more closely related subgroups of genes (Joseph et al., 2008). A comparison of the actin gene

family in *D. discoideum* with that of *Dictyostelium fasciculatum*, a member of the more primitive group 1 clade of dictyostelids (Romeralo et al., 2011), suggests that while both organisms have large actin gene families their expansion occurred independently in evolution (Joseph et al., 2008). This further supports the idea that a large actin gene family is a feature of amoeboid genomes.

1.5.2 Regulation of *Dictyostelium* actin genes

It has long been known that *D. discoideum* contains many different actin genes (McKeown et al., 1978; Kindle & Firtel, 1978). Despite encoding multiple copies of the gene, a single protein isoform was shown to comprise around 95% of the total actin content of cells (Vandekerckhove & Weber, 1980). Therefore, Act8 proteins appear to be the major isoform of actin in *D. discoideum*, and probably comprise the majority of actin-based structures in the cell, given that overexpression of another isoform, Act3, alters the normal actin network (Joseph et al., 2008). Actin was shown to make up around 8-10% of the total protein content of vegetative *D. discoideum* cells (Uyemura et al., 1978), with relative synthesis of the protein increasing dramatically in early development before decreasing to below vegetative levels at later time points (Tuchman et al., 1974; Alton & Lodish, 1977; Margolskee & Lodish, 1980). This change in the protein level was concomitant with changes in total actin mRNA levels (Margolskee & Lodish, 1980; Romans et al., 1985). Furthermore, actin gene expression was shown to be biased towards pre-stalk and stalk cells in the developing organism (Alton & Brenner, 1979; Coloma & Lodish, 1981; Tsang et al., 1982; Mehdy et al., 1983; Barklis & Lodish, 1985).

Subsequent studies attempted to decipher the regulatory control mechanisms for different actin genes. Sequencing of the genes showed that despite the highly similar coding regions only limited sequence similarity existed in the highly A/T-rich flanking regions (Firtel et al., 1979; McKeown & Firtel, 1981a,b; Kimmel & Firtel, 1983). Two sizes of actin mRNA were also

identified (Kindle & Firtel, 1978; Romans & Firtel, 1985b). More extensive sequencing efforts on a greater number actin genes identified conserved motifs in the upstream regions of the gene family (Romans & Firtel, 1985b) potentially involved in differential control of the gene family. The same authors also performed nuclease protection assays using the more diverse 5' UTRs to determine specific developmental expression profiles for different family members (Romans et al., 1985). These experiments probed very short, low complexity sequences in actin transcripts and therefore results are potentially conflicted by cross-hybridisation. However, the authors identified some different expression patterns, with *act8* mRNA constantly expressed at high levels throughout early and mid development, while *act6* and *act7* showed a sharp increase in expression at 2.5 h which quickly decreased by 5 h of development (Romans et al., 1985).

Having identified the potential for differential control of gene expression during development by certain regulatory motifs, later studies attempted to determine the precise sequences required for actin gene expression. By fusing promoter sequences to a selectable marker, Knecht et al. (1986) showed that sequences upstream of the *act6* and *act15* genes were sufficient to drive developmental expression patterns. Further studies identified a specific palindromic sequence motif required for expression of these genes (Cohen et al., 1986; Nellen et al., 1986). This was not found in *act8* which exhibited a different developmental expression pattern. Hori & Firtel (1994) showed that this sequence motif was sufficient to drive growth phase expression of a developmental gene. They also showed that the length of poly(dT) tracts common to all actin gene promoters (Romans & Firtel, 1985b) scaled with expression level, which was in agreement with other data suggesting that such tracts led to reduced nucleosome occupancy at the promoter (Struhl, 1985). Since then, excepting work by Joseph et al. (2008), very little has been done to develop our understanding of the actin gene family, as opposed to the protein and its functions, in *Dictyostelium*.

1.6 Aims of the thesis

In this study I have made use of recently available techniques to explore the factors influencing the evolution of the highly expanded actin gene family in *Dictyostelium discoideum*, and more specifically the group of 17 genes encoding an identical protein. It is commonly assumed that expansion of highly similar gene families such as those of rDNA and histone genes simply enables increased protein production for these already strongly expressed genes. Indeed, histone genes are thought to be under 'coordinate control' throughout the cell cycle (Holmes et al., 2005). However, not all histones have the same promoter regulatory elements (Chowdhary et al., 2005) and this might explain differential contributions to overall histone content in different cell types (Holmes et al., 2005). With studies from the 1980s showing variable actin gene expression patterns, I have used multiple different techniques to determine whether differential regulation, as well as high levels of protein production, might be a contributory factor in the evolution of an expanded, highly expressed gene family.

Chapter 2

Materials and methods

2.1 Molecular biology

2.1.1 Bacterial strains

Routine subcloning and plasmid preparation was done using the *E. coli* strain, Top Ten [F^- *mcr*A Δ(*mrr-hsdRMS-mcrBC*) Φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*U *gal*K *rps*L (Str^R) *end*A1 *nup*G] (Thermo Fisher Scientific).

2.1.2 Bacterial growth

Bacteria were grown in shaking cultures in Luria-Bertani (LB) medium and on LB agar plates at 37°C. Selection with 100 μg/ ml ampicillin (Sigma) was used where appropriate.

2.1.3 Transformation of bacteria

Competent bacterial cells were transformed with DNA ligation reaction products as follows. 100 μl competent cells were added to a 10 μl ligation reaction. This mixture was incubated on ice for 10 minutes before heat shock at 42°C for 1 minute with subsequent incubation on ice for 2 minutes. Transformed cells were plated directly on LB agar plates containing 100 μg/ml ampicillin and grown at 37°C overnight.

2.1.4 Plasmid preparation

Bacterial cultures for plasmid extraction and purification were grown at 37 °C overnight with shaking. Small scale plasmid preparation for identifying recombinant transformants and DNA sequencing was done using 2 ml LB cultures. Bacteria were lysed and plasmid DNA extracted using QIAprep Spin Miniprep Kit (QIAGEN).

Large scale plasmid preparation for transformation of *Dictyostelium* cells was done using 150-200 ml LB cultures. Plasmid preparation was done using QIAGEN Plasmid Maxi Kit (QIAGEN). DNA concentrations were measured using a NanoDrop 1000 (Thermo Fisher Scientific).

2.1.5 DNA digestion

For routine subcloning 500 ng-2 μg DNA was digested according to New England Biolabs (NEB) protocols. For reactions containing a single restriction endonuclease, calf intestinal alkaline phosphatase (CIP; NEB) was added to digestions containing vector backbone to remove 5' phosphate groups and prevent self-ligation. Filling in of 3' recessed restriction sites was done using DNA Polymerase I, Large (Klenow) Fragment (NEB) in the presence of 33 μM dNTPs.

For digestion of DNA for transformation of *Dictyostelium* cells, where amounts of DNA to be digested ranged from 10-100 μg, NEB protocols were followed as before but incubation times were increased to prevent excessive use of restriction enzymes.

2.1.6 Agarose gel electrophoresis of DNA

10x agarose gel loading buffer (40% sucrose, 0.4% bromophenol blue) was added to DNA samples for electrophoresis through 1% agarose gels (1% molecular biology grade agarose (Eurogentec)) and 300 ng/ml ethidium bromide in Tris acetate buffer (TAE: 40 mM Tris-acetate, 1 mM EDTA) at 100-135 V (Mupid One electrophoresis apparatus, Takara). Samples were run alongside a DNA ladder (1 kb DNA ladder, NEB) with DNA fragments vi-

sualised by UV transillumination. DNA fragments for recombinant plasmid construction were purified following electrophoresis using QIAquick Gel Extraction Kit (QIAGEN).

2.1.7 DNA ligation

Ligation of vector and insert DNA was performed using 1 unit of T4 DNA ligase (Promega) and approximately 20 ng vector DNA and 50 ng insert DNA in a 10 μl reaction volume. Reactions were incubated at 25°C for 30 minutes before transformation into competent *E. coli* cells.

2.1.8 DNA constructs

2.1.8.1 MS2 and PP7 knock-ins

Promoter and coding sequence homology arms were amplified by PCR with the following primers (restriction sites highlighted in bold; F and R denote forward and reverse primers):

Promoters

act1 F1: 5' CGAGA**ATCGAT**TTGTAATCGTGTTTGGGT 3'

act1/act6 R1: 5' CGAGA**TGTACA**AAAGCTTGAACATCTTCACCATCCA 3'

act6 F1: 5' CGAGA**ATCGAT**CAAAGTTGCGCAAAATAATA 3'

act8 F1: 5' CCAGA**ATCGAT**CTTAACTTTTCTTTATGAAAGAG 3'

act8 R1: 5' CGAGA**TGTACA**AAAGCTTGAACATCTTCACCGTCCA 3'

Coding sequences

act (all) F2: 5' CGAGA**ACTAGT**GATCCTGGTCGTCCAAGACACACTG 3'

act1 R2: 5' CGAGA**GCGGCCGC**TTGACTTGTTCCTGTAAA 3'

act6 R2: 5' CGAGA**GCGGCCGC**TTGATGGCTGGATAAA 3'

act8 R2: 5' CGAGA**GCGGCCGC**GATTTTCATACTCTTTGGC 3'

A further primer common to all three genes was then used to amplify the final coding sequence homology arms from initial cloned PCR products due to 3' sequence errors:

act R3: 5' CGAGA**GCGGCCGC**CTGGGAACATAGTTGTACCA 3'

Amplification of promoter sequences utilised primers containing restriction endonuclease cut sites for ClaI and BsrGI, while gene sequences were amplified with SpeI and NotI sites. These sequences were cloned into pBluescript II plasmids containing a modified multiple cloning site (MCS) with a BsrGI-SpeI linker (GGATC**TGTACA**GAA**ACTAGT**GATCC) inserted into a (subsequently destroyed) BamHI site. These sequences were checked by Sanger sequencing against the Ax4 genomic reference sequences downloaded from dictybase.org. Following correct assembly of both homology arms, a DNA fragment consisting of a cassette of 24 MS2 or PP7 bacteriophage-derived hairpin loop sequences coupled to a blasticidin resistance gene (*bsr*) sourced from in-house laboratory plasmids as described previously (Corrigan et al., 2016) was inserted between the homology arms using BsrGI and SpeI restriction enzymes (see figure 2.1 A). LoxP sites flanking the *bsr* gene enabled its removal after selection of positive clones.

Care was taken during the cloning process to ensure the maintenance of the full complement of repeat sequences as these are readily susceptible to recombination-mediated expansion or deletion. Bacterial cultures harbouring MS2- or PP7-containing plasmids were grown for as little time as necessary to achieve log phase growth due to the fact that overgrown cultures display increased frequency of adverse recombination events. This effect was mitigated further by growing freshly transformed bacteria on agar plates at 22 °C as opposed to 37 °C. Before transformation of constructs into *Dictyostelium* cells plasmid preparations were assessed for MS2 repeat length by digestion with BsrGI/EcoRI before running on an agarose gel. A full complement of MS2 repeats should run at about 1.35 kb. Correct integration of the construct was confirmed by Southern blots (Fig. 2.1).

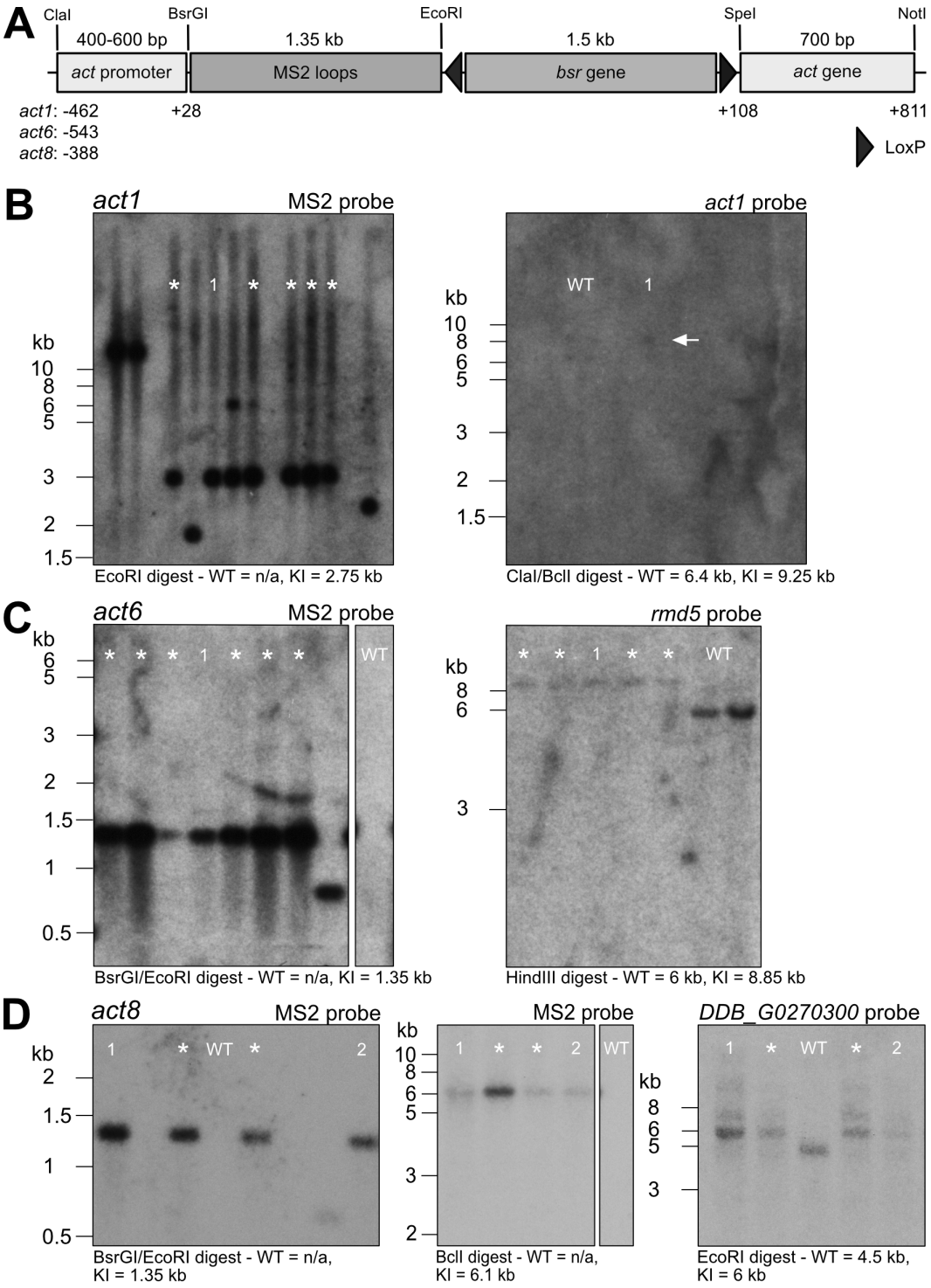

Figure 2.1: *act*-MS2 targeting construct and Southern blot determination of positive clones. (A) Construct for targeting MS2 loops and *bsr* selection marker to three different actin genes. (B)-(D) Southern blots for each actin gene confirming correct targeting on the construct. Asterisks indicate positive clones, (1) and (2) indicate independent clones used in subsequent experiments. Ax3 (WT) genomic DNA was used as a control. Multiple blots were done for each gene to check the fidelity of the MS2 repeats (1.35 kb in length) as well as ensuring targeting to the correct locus. See figure 2.2 for a schematic of these loci.

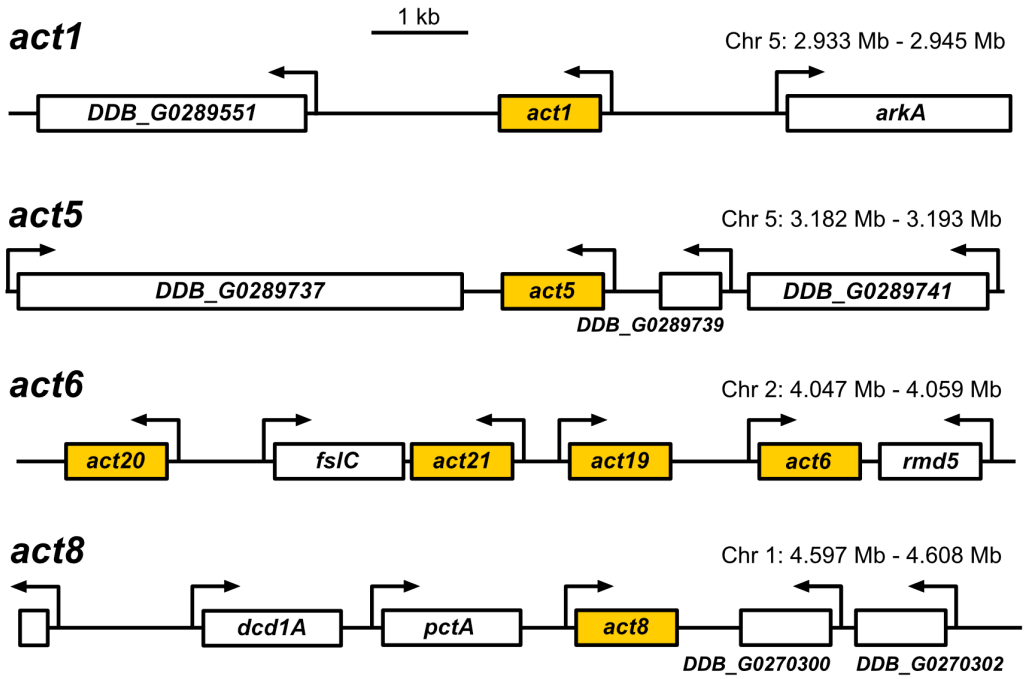

Figure 2.2: Genomic loci of actin genes studied in this thesis. Provided as a point of reference for construct generation and Southern blots.

2.1.8.2 Promoter switches

Knock-in constructs were generated to 'switch' the promoters of *act6* and *act8* genes at their endogenous loci, largely using existing reagents detailed above (see Fig. 2.3). The SpeI/NotI coding sequence fragments of the *act6* and *act8* genes were spliced next to the opposite promoter as shown in figure 2.3. A region upstream of the promoters of each gene was then amplified as the 5' homology arm for the construct. To target the *act8* promoter to the *act6* locus a sequence containing a 1 kb 3' portion of the neighbouring *act19* coding sequence and 3' UTR was amplified using the following primers:

act19 F1: 5' CGAGA**GGGCCC**CCAGACGGTCAAGTTATCACA 3'

act19 R1: 5' CGAGA**GGGCCC**CCATCATTATTCTTCAA 3'

To target the *act6* promoter to the *act8* locus a sequence containing 900 bp of the neighbouring *pctA* gene and 3' UTR (see figures 2.2, 2.3) was amplified using the following primers:

pctA F1: 5' CGAGA**GGGCCC**GTTGGTGCAGATGGTAAA 3'

pctA R1: 5' CGAGA**GGGCCC**CGATTTGATGGGTCACTTAA 3'

These PCR products were then cloned into the switched promoter-coding sequence vectors using the upstream ApaI restriction site with the orientiation checked by EcoRI. Finally, an MS2-bsr cassette (as described above) was inserted using BsrGI and SpeI restriction enzymes. All sequences were checked by Sanger sequencing before transformation.

2.1.8.3 NeonGreen fusion proteins

Gene-specific homology arms for *act5*, *act6* and *act8* were used to construct vectors to target the fast-folding yellow fluorescent protein variant mNeonGreen (Shaner et al., 2013) to the 3' end of actin gene loci (Fig. 2.4). SpeI/NotI fragments of actin gene sequences (cloned by PCR using *act* F2 and gene-specific R2 primers, see 2.1.8.1) were spliced into a pBluescript II vector with the BamHI site filled in by Large Klenow fragment (NEB). Coding sequences from the 3' end of each gene were synthesised (Eurofins) from the common BspHI site (+848) to the penultimate codon (+1128) followed by a GlyGlyProPro linker sequence and BamHI and NotI restriction sites. The synthesised products were cloned into the gene sequence vectors using BspHI and NotI. Gene-specific 3' UTR sequences were cloned in using BamHI and NotI and the following primers:

act (all) F3: 5' CGAGA**GGATCC**CCACAGAAAATGTTTCTAA 3'

act5 R3: 5' CGAGA**GCGGCCGC**CTATGATTGCTCTTTGTAA 3'

act6 R3: 5' CGAGA**GCGGCCGC**GTGTATTGAAGGGTGTAA 3'

act8 R3: 5' CGAGA**GCGGCCGC**GTACAAGACGGCTTGAAT 3'

Finally, a cassette containing *mNeonGreen* coding sequence (adapted to *Dictyostelium* codon bias using gene synthesis by Eurofins) coupled to *bsr*

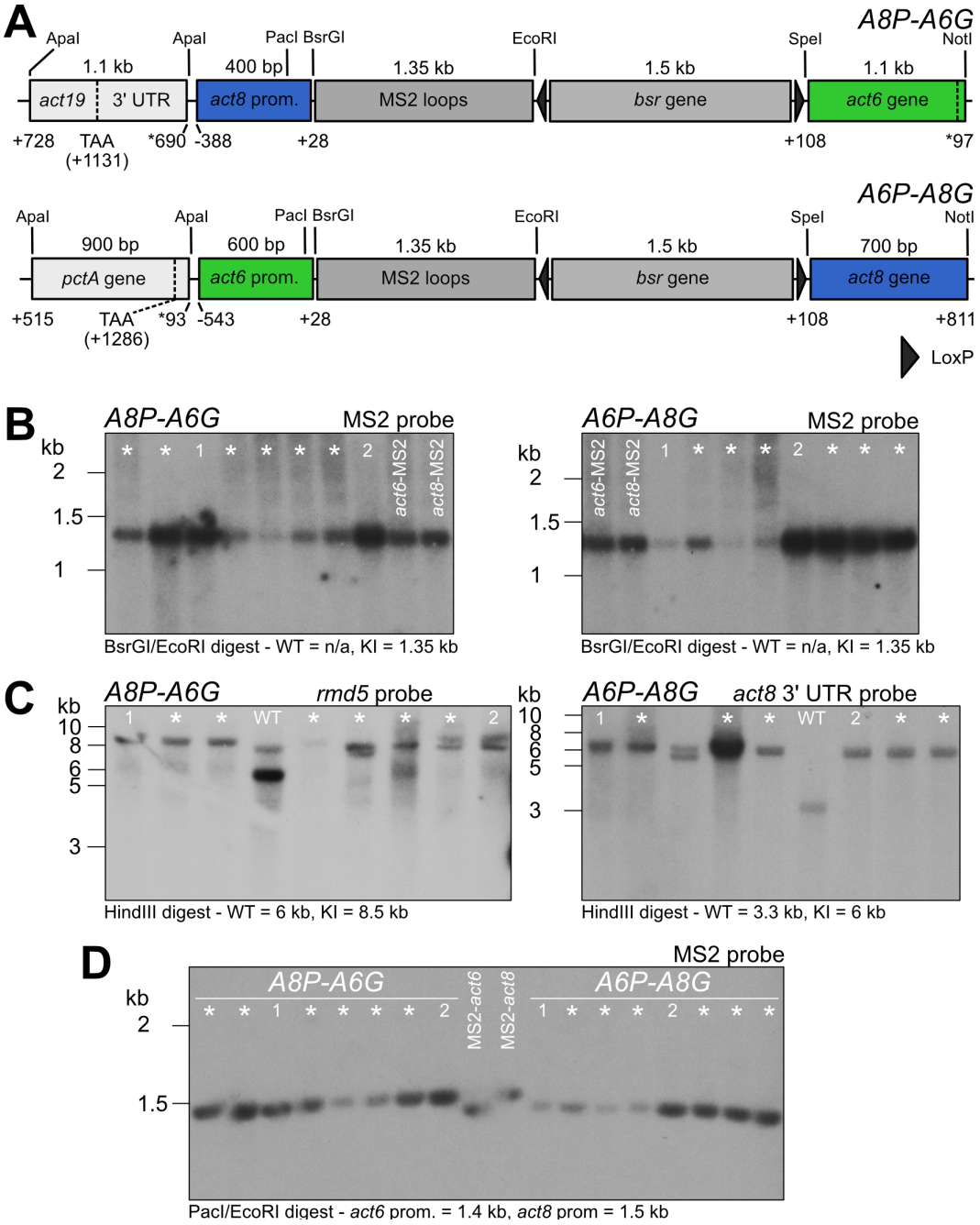

Figure 2.3: Targeting constructs and Southern blots for switching actin promoters. (A) Two targeting constructs were used to target the *act8* promoter to the *act6* locus and *vice versa*. In each case, a region around 1 kb was inserted into the vector upstream of the promoter to ensure proper targeting to the appropriate locus. (B-C) Multiple Southern blots were done for each cell line to ensure MS2 repeat length fidelity and correct targeting. Ax3 (WT) genomic DNA was used as a control. (D) A final blot was done confirming the presence of the opposite promoter at each locus by taking advantage of differentially positioned PacI sites in the two promoters (see A). *act6*-MS2 and *act8*-MS2 (see figure 2.1) were used as controls. Asterisks indicate positive clones, (1) and (2) indicate independent clones used in subsequent experiments. See figure 2.2 for a schematic of these loci.

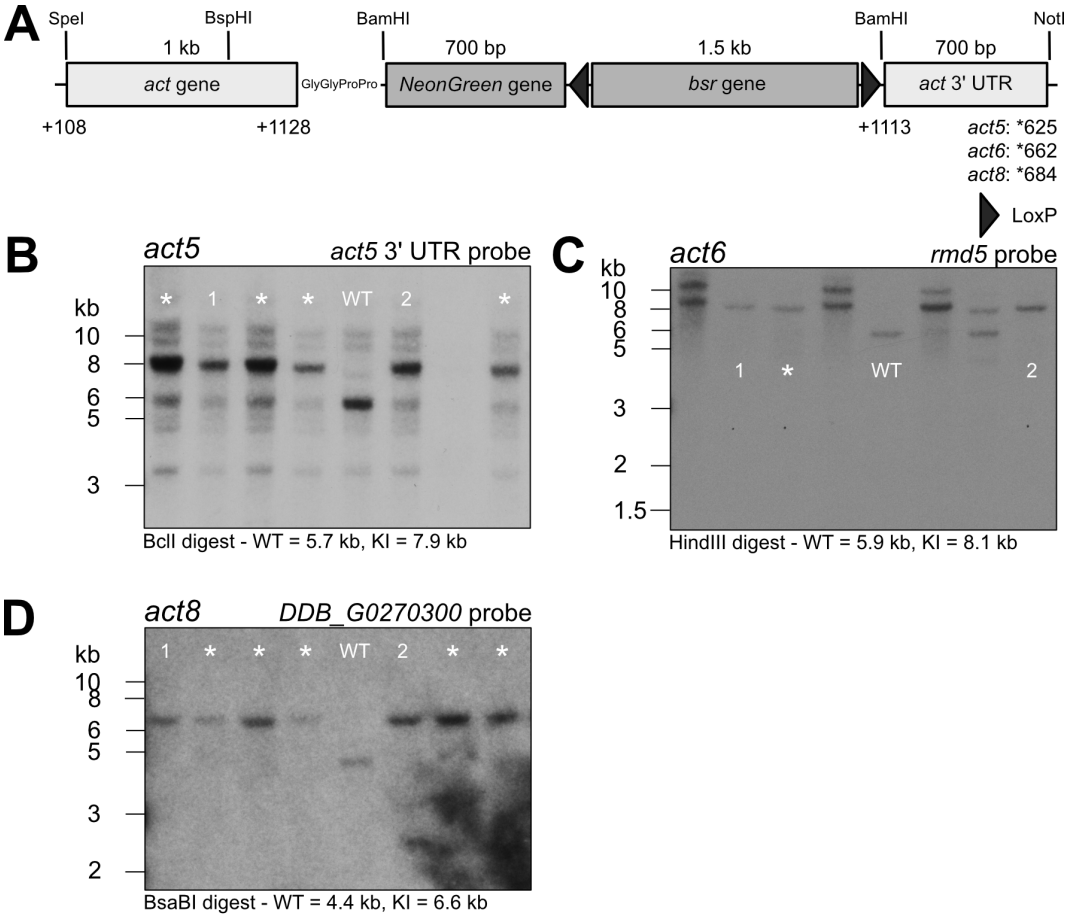

Figure 2.4: Actin-mNeonGreen targeting construct and clone identification by Southern blot. (A) Actin-mNeonGreen fusion proteins were generated by targeting the *mNeonGreen* gene to individual actin gene loci using the coding sequence and 3' UTR as targeting arms. (B-D) A single Southern blot was required to confirm proper targeting of positive clones (asterisks). (1) and (2) indicate independent clones used in subsequent experiments. Ax3 (WT) genomic DNA was used as a control. See figure 2.2 for a schematic of these loci.

was cloned in using the BamHI site. Correct directionality of the NeonGreen-bsr cassette was checked for using an EcoRV/NotI digest. LoxP sites flanking the *bsr* gene enable removal after selection of positive clones.

2.1.8.4 Actin gene knockouts

To knock-out a significant proportion of the gene family the cluster of four actin genes on chromosome 2 was targeted for deletion (see figure 5.1 for schematic). A 1 kb 5' homology arm of *act20* 3' UTR and a portion of coding sequence was cloned into pBluescript II using HindIII and EcoRI and the

following primers:

act20 F1: 5' CGAGA**GAATTC**CCAGACGGTCAAGTTATCACA 3'

act20 R1: 5' CGAGA**AAGCTT**TTACAATATCAACTCTCA 3'

A 3' homology arm of similar length corresponding to a 3' portion of the *rmd5* coding sequence and downstream 3' UTR was then cloned into this vector using EcoRI and NotI sites and the following primers:

rmd5 F1: 5' CGAGA**GCGGCCGC**GGTGATATATTTGCAAATG 3'

rmd5 R1: 5' CGAGA**GAATTC**CCCAAAATTCCTTTATTGTA 3'

Finally, after ensuring the correct sequences of these homology arms by Sanger sequencing the *bsr* gene was spliced into the EcoRI site, with directionality checked by EcoRV. Knocking this construct into cells inactivates each of the four actin genes in the cluster but also deletes the coding sequence of another gene, *fslC* situated between *act20* and *act21*. Therefore, to allow us to control for the effect of deleting this gene we also generated an *fslC* knockout construct using the following primers:

fslC F1: 5' CGAGA**AAGCTT**GATAGGCAGTGAAAAAGTT 3'

fslC R1: 5' CGAGA**GAATTC**CTAATGGGCTTGGACAAA 3'

fslC F2: 5' CGAGA**GAATTC**GAGTTTAATGGATATTGT 3'

fslC R2: 5' CGAGA**GCGGCCGC**CTATTAGTATTTGAATCTT 3'

fslC F1 and R1 primers were used to generate the 5' homology arm, with the F2 and R2 primers producing the 3' homology arm by PCR. Each arm was individually cloned into pBluescript II using HindIII and EcoRI (5' arm) and EcoRI and NotI (3' arm). The *bsr* gene was then spliced into the EcoRI restriction site.

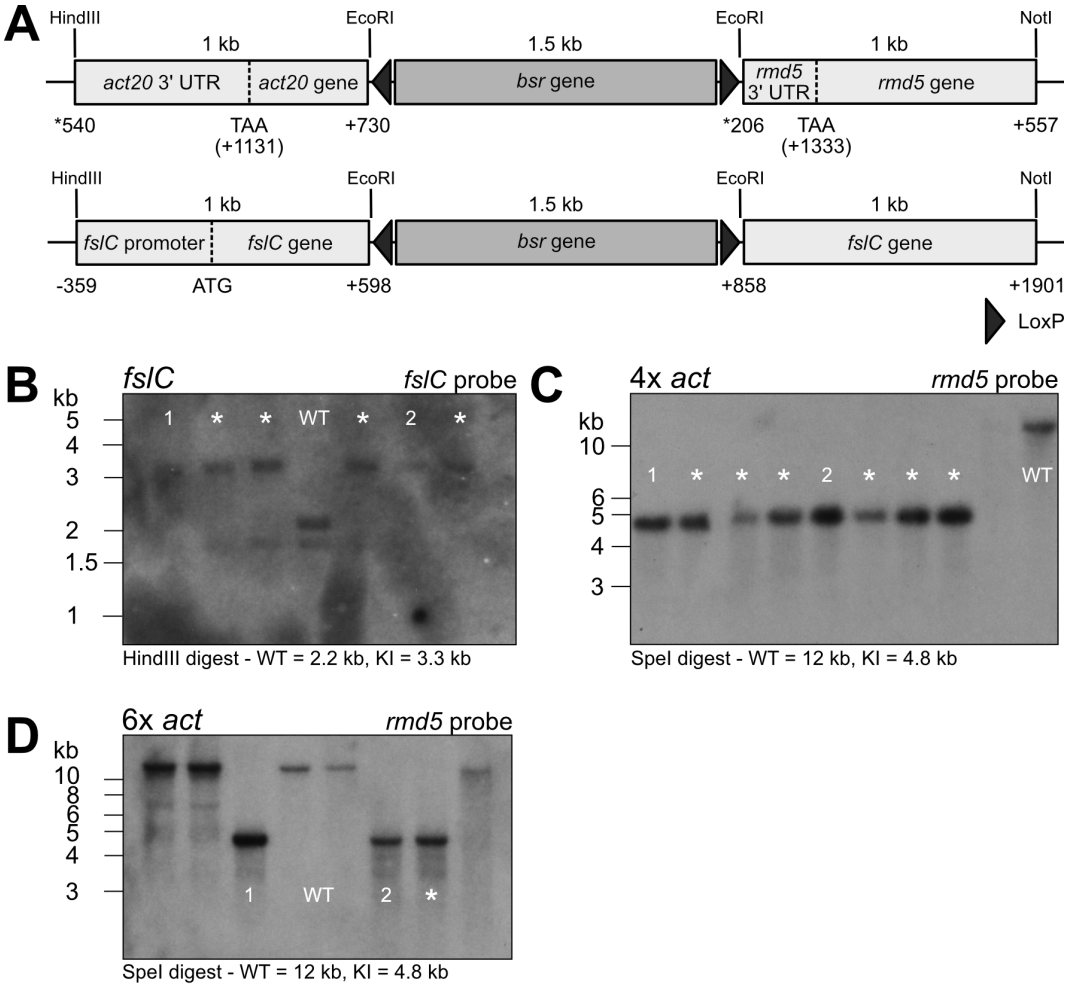

Figure 2.5: Targeting constructs and Southern blot of confirmation of ΔfslC and actin gene knockout cell lines. (A) Two targeting constructs to delete four actin genes (top) and disrupt the *fslC* gene (bottom) were generated as described in text. (B-C) Southern blot confirmation of positive clones (asterisks) for each construct knocked into Ax3 cells. (D) Southern blot confirmation of positive clones (asterisks) for actin gene knockout construct transformed into a cell line with two actin genes disrupted by MS2 and PP7 constructs previously. A total of six actin genes are deleted in these strain. (1) and (2) indicate independent clones used in subsequent experiments. Ax3 (WT) genomic DNA was used as a control. See figure 2.2 for a schematic of these loci.

2.2 *Dictyostelium* cell culture

2.2.1 Strains

Dictyostelium discoideum Ax3 cells were used for all experiments described.

2.2.2 Culture conditions

Dictyostelium cells were maintained axenically in HL5 medium (including glucose; Formedium) at 22°C. For co-culture with bacteria, 100 μl of a cell suspension of *Klebsiella aerogenes* in LB was added to SM agar plates (1L: 10 g glucose, 10 g proteose peptone, 5 g yeast extract, 1 g $MgSO_4 \cdot 7H_2O$, 1.9 g KH_2PO_4, 0.6 g K_2HPO_4, 20 g bacto agar) along with 10^5-10^6 *Dictyostelium* cells. Plates were incubated in a humidified atmosphere at 22°C.

For induction of development, cells were harvested into KK2 (20 mM KPO_4 pH 6.2) and washed several times to remove all food sources (bacteria or HL5 medium). 3×10^6 cells were then plated on 1.5% 35 mm KK2 agar plates in 1-2 ml of KK2 and allowed to settle, before removal of KK2 and placement in a humidified chamber at 22°C.

2.2.3 Transformation and selection

Cells were grown axenically in HL5 on plastic petri dishes until 50-80% confluency before being transferred to a shaking culture in HL5 medium. Upon reaching log phase growth (1-5$\times10^6$ cells/ml) approximately 5×10^6 cells were harvested and washed once in ice-cold H-50 buffer (20 mM HEPES buffer pH 7.0, 50 mM KCl, 10 mM NaCl, 1 mM $MgSO_4 \cdot 7H_2O$, 5 mM $NaHCO_3$, 1 mM $NaH_2PO_4 \cdot H_2O$) before resuspension in 100 μl H-50 buffer. This cell suspension was then added to approximately 4 μg of linearised DNA and gently mixed by pipetting before being transferred to a pre-cooled 1 mm cuvette (Molecular Bioproducts) and incubated on ice for 5 minutes. Cells were electroporated using a Bio-Rad Gene Pulser XCell system with an exponential decay protocol at 750 V, 25 μF. Following electroporation, cells were immediately plated in HL5 for recovery, with selection added after approximately 16 h. Transformants were selected using either 10 μg/ml blasticidin S (Calbiochem) or 20 μg/ml G418 (Thermo Fisher Scientific) dependent on the construct being transformed. Selection was maintained for 10 days in blasticidin and 3-5 days in G418. Blasticidin transformants were subsequently single-cell cloned out onto SM plates with lawns of *Klebsiella*,

while G418 transformants were grown up in HL5 medium. Two independent clones were generated (and used equally in all experiments) for all genetically modified cell lines by independent transformations with the exception of *act1*-MS2 and *act6*-MS2 cell lines.

Prior to imaging, cells with an integrated MS2-*bsr* construct had the *bsr* gene removed by Cre-Lox recombination to enable natural termination of transcription. Cells were transformed as described above with a pDEX-Cre-NLS plasmid for transient expression of Cre recombinase. Single clones were then replica-plated in 96-well plates and grown with HL5 in the presence or absence of blasticidin S. Blasticidin-sensitive cells were transferred to 6-well plates for expansion and confirmation of sensitivity.

2.2.4 Cryopreservation

Cells were cryopreserved by freezing in a solution of HL5:FBS:DMSO (47.5:47.5:5) at –80 °C or in liquid nitrogen. Upon thawing, cells were either placed in HL5 on petri dishes, with any required selection added after 16 h, or a small amount of frozen material added to a SM plate with a pre-spread cell suspension of *Klebsiella*.

2.3 Analysis of cellular DNA, RNA and protein

2.3.1 Radio-labelled probes

Radioactive probes for Northern and Southern blotting were generated using a High Prime DNA labelling kit (Roche, Basel, Switzerland). A random priming reaction was performed with radioactive dCTP [α-^{32}P] (PerkinElmer) according to manufacturer's instructions. Probes were generated by PCR and subsequent subcloning into pBluescript II or digestion of existing vectors (see figures 2.1, 2.3, 2.4, 2.5, 4.22, 5.2 for details).

2.3.2 Extraction of *Dictyostelium* DNA

10^8-3×10^8 bacterially grown cells were harvested, washed once in KK2 to remove bacteria then resuspended in 20 ml KK2 and rocked for 2 hours to

allow for cellular protein degradation. Cells were pelleted at 2000 rpm for 2 min and taken up in 1 ml ice-cold RLB (0.32 M sucrose, 10 mM Tris-HCl pH 7.5, 5 mM $MgCl_2$, 1% Triton X-100) for lysis while kept on ice. Nuclei were then pelleted at 5000 rpm for 5 min and resuspended in 50 μl RLB (on ice). 350 μl GPA (10 mM Tris-HCl pH 7.5, 10 mM EDTA) and 350 μl GPB (10 mM Tris-HCl pH 7.5, 0.7% SDS) were then added, followed by 60 μg Proteinase K (Ambion). Samples were mixed by inversion and incubated at 57°C for 4-16 h. Any remaining protein was removed by phenol extraction (add equal sample volume of phenol, mix and centrifuge 13,000 rpm for 2 min, pipette off top phase, repeat once more). DNA was precipitated in 2 volumes 100% ethanol and 0.1 volumes 3M NaAc pH 5.2, washed in 70% ethanol and resuspended in TE (10 mM Tris-HCl, 1 mM EDTA).

2.3.3 Southern blot analysis of *Dictyostelium* DNA

Genomic DNA was digested overnight (see figures 2.1, 2.3, 2.4, 2.5 for details) before electrophoresis on 0.8% agarose gels. DNA was then depurinated by rocking gels in 0.125 M HCl for 10 min, denatured in 0.5 M NaOH/1.5 M NaCl for 30 min and neutralised in 0.5 M Tris-HCl/1.5M NaCl pH 7.5 for 30 min. DNA was transferred to positively charged nylon membrane (Hybond N+, GE Healthcare, Chicago, IL) by capillary transfer in 5x SSC (750 mM NaCl, 75 mM $Na_3C_6H_5O_7 \cdot 2H_2O$) overnight. DNA was crosslinked to the blot by UV irradiation (UV Stratalinker, Stratagene) before pre-hybridisation in 30 ml Church phosphate buffer (0.5 M $NaPO_4$, 7% SDS, 10 mM EDTA, pH 7.2) for 1-2 h in a hybridising oven (HB-1D hybridiser, Techne) at 65°C. Hybridisation with denatured probe (5 min at 100°C) was then added and hybridised for 4-16 h. The blot was then washed with 1x SSC/0.1% SDS for 15 min and 0.1x SSC/0.1% SDS for 15 min before visualisation by autoradiography.

2.3.4 Extraction of *Dictyostelium* RNA

Pellets of 10^7 cells were collected, centrifuged and supernatant aspirated before immediately being placed at –80 °C. After all samples were obtained, pellets were thawed on ice and resuspended in 500 μl SDS lysis buffer (50 mM Tris-HCl pH 7.5, 0.1 M NaCl, 10 mM EDTA, 1% SDS). 500 μl phenol was added and samples were placed on a rotating wheel for 30 min. Mixed samples were centrifuged at 13,000 rpm for 3 min before collection of the top phase and addition of 1 ml ethanol (abs.). Following precipitation at –20 °C for 20 min samples were washed in 70% ethanol and resuspended in MilliQ H_2O.

2.3.5 Northern blot analysis of *Dictyostelium* RNA

RNA concentration of samples was determined using a NanoDrop 1000 (Thermo Fisher Scientific). 5-10 μg of RNA was mixed in equal volumes with a loading cocktail (75% formamide, 7.5% formalin, 17.5% agarose gel loading buffer) before loading on a 1% agarose, 0.5% formaldehyde gel made with, and electrophoresed in, MOPS buffer (20 mM MOPS pH 7, 5 mM NaAc, 1 mM EDTA, final pH 7). Capillary transfer, cross-linking, hybridisation, probe preparation and visualisation were performed the same as for Southern blot analysis, omitting the depurination, denaturation and neutralisation washes before transfer.

2.3.6 Analysis of *Dictyostelium* mRNA stability

Cells were seeded onto 60 mm petri dishes and incubated at 22 °C in HL5 medium for 16 h to arrive at a final confluence of 40-60%. Cells were washed with HL5 2 h before addition of HL5 containing actinomycin D to a final concentration of 125 μg/ml (Muramoto et al., 2012). Cells were harvested at 0, 15, 30, 60 and 90 min after addition of actinomycin D, centrifuged (2000 rpm, 2 min) and medium removed before snap freezing in dry ice. A control time-course was done with DMSO alone. RNA extraction and Northern analysis was done as described above.

2.4 Flow cytometry

Fluorescence levels of actin fusion protein expression in individual cells was determined by flow cytometry using a LSRII Flow Cytometer (Beckton Dickinson). Cells were grown on petri dishes in HL5 medium for four days at mid-log phase prior to the experiment. Before using the flow cytometer, cells were blasted off the petri dish and washed twice in KK2 to remove autofluorescent medium. For each cell line, including Ax3 as a control, 10,000 cells were assessed. Identical gates were outlined for cell lines monitored on the same day to remove debris and clumped cells. Visualisation of fluorescence distributions in the FITC channel was done using FlowJo software (FlowJo).

2.5 Microscopy

2.5.1 Fluorescence microscopy

2.5.1.1 Live cell transcription dynamics

In order to visualise transcription at a particular gene locus a plasmid containing the specific MS2 coat protein (MCP) fused to GFP (as described (Chubb et al., 2006)) was transformed into cells containing an MS2-tagged gene (Fig. 2.1). Before imaging, cells were maintained in mid-log phase growth for 4-5 days with G418 selection in HL5. 16 h prior to imaging cells were transferred to an imaging dish (2- or 4-well Nunc Lab-Tek II Chambered Coverglass, ThermoFisher Scientific) in imaging medium (70% LoFlo medium (Formedium), 20% HL5, 10% FBS) without selection. 2 h prior to imaging the medium was changed.

Cells were imaged on an UltraVIEW VoX spinning disc confocal microscope (TiE inverted stand, Nikon; CSU-X1 spinning disc scanning head, Yokogawa) with an EMCCD camera (C9100-13, Hamamatsu) and a 60x objective lens. 488 nm (7% laser power) and 561 nm (3%) DPSS lasers were used to visualise MCP-GFP and H2B-mCherry respectively with 70 ms exposures. For each XY position an 11μm Z-stack was taken with 0.45 μm

Z-slices every 45 s for 30-90 min.

2.5.1.2 RNA fluorescence *in situ* hybridisation

Cells were maintained in identical culturing conditions to those described above in terms of medium, selection and imaging dishes. Cells were then fixed for 10 min with a 2× fixation solution (1× PBS, 7.4% formaldehyde, final volume with nuclease free H_2O (Invitrogen)) by adding 1 ml to each well containing 1 ml of imaging medium. Samples were washed twice with 1× PBS with 5 min incubation before permeabilisation with 70% ethanol for 1 h at room temperature. Pre-hybridisation was then done by incubating samples with 1 ml wash buffer (2× SSC, 10% formamide, final volume with nuclease free H_2O (all Invitrogen)) for 5 min. 150 μl of hybridisation buffer (50 nM Stellaris probe (Biosearch Technologies), 10% dextran sulfate, 2× SSC, 10% formamide, final volume with nuclease free H_2O (all Invitrogen)). An 18×18 mm coverslip was then used to spread the hybridisation buffer over the entire sample. Samples were then placed in a petri dish with a wet towel and sealed with parafilm to ensure a humid environment and incubated at 37°C overnight. 1 ml of wash buffer was added to remove coverslips before two further wash steps at 37°C for 30 min. After removal of wash buffer, 1 ml of 1× DAPI solution in 2× SSC was added and incubated at room temperature for 5 min. Immediately before imaging samples were washed in 1× PBS before incubation in 1 ml GLOX buffer (2× SSC, 0.4% glucose, 10 mM Tris-HCl) for 1-2 min. After equilibration, GLOX buffer was removed and replaced with 150 μl GLOX buffer containing 1% glucose oxidase stock solution (15% glucose oxidase, 50 mM NaAc pH 5.2, final volume with nuclease free H_2O) and 1% catalase (Sigma-Aldrich), again spread by a coverslip. GLOX buffer with enzymes was changed every 2-3 hours if necessary, to prevent loss of anti-fade activity.

Cells were imaged on an UltraVIEW VoX as described, but with 100× objective lens. Far-red probes were imaged with a 640 nm laser at 15% power with 5 s exposure. DAPI was imaged using a 405 nm laser at 5.5%

power with 400 ms - 1 s exposure. For each image, a Z-stack of between 10-14 μm with 0.25 μm slices was taken.

2.5.2 Brightfield microscopy

In order to image random cell motility cells were harvested from pre-clearing agar plates (SM plates where the lawn of bacteria has not been cleared, such that none of the cells have initiated the starvation response). Two protocols to isolate *Dictyostelium* cells were used before imaging (described further in section 5.5), one more stringent than the other.

Normal protocol Two pre-clearing plates were harvested and washed twice in 20 ml KK2 with centrifugation at 2000 rpm for 2 min. Cells were then allowed to settle on a plastic 6-well dish for 15 min before washing by aspiration a further three times with KK2, taking care not to blast off attached *Dictyostelium* cells.

Stringent protocol As before two pre-clearing plates were harvested and washed with KK2. Then, again, cells were allowed to settle on 6-well dishes before washing by aspiration with KK2. Cells were then removed from the bottom of the wells by blasting off and subsequently washed three times by centrifugation (by pulsing a microcentrifuge to 13000 rpm). Finally, cells were again allowed to settle for 7 min and washed three more times with KK2 by aspiration.

Cells were imaged in 6-well dishes on an Observer Z1 microscope (Zeiss) with 12-bit Orca ER camera (Hamamatsu) and a 10$\times$ phase contrast objective lens. A single image was captured with 40 ms exposure for each of 4-8 fields of view every 30 s.

2.6 Image analysis

2.6.1 Live cell transcription dynamics

Analysis of transcription spot intensities over time in individual cells was done using a custom MATLAB pipeline (written by Adam Corrigan) as described in (Corrigan et al., 2016). Briefly, cells and nuclei were segmented

before being collected into single cell tracks. Features including spot intensity, xy coordinates, background fluorescence intensity and local density were extracted. Spot intensity traces were corrected for the average nuclear background GFP intensity for each cell. Incorrectly assigned tracks were manually corrected and any tracks containing multi-nucleate cells removed.

2.6.2 RNA fluorescence *in situ* hybridisation

RNA FISH images were analysed using the FISH-quant MATLAB package (Mueller et al., 2013). Various programmes for quantification of FISH images were tested by lab members and FISH-quant was deemed the most accurate and robust. FISH-quant allows the extraction of the number of mRNA transcripts in both cytoplasm and nucleus as well an estimate of the number of nascent transcripts at the transcription site of the gene of interest.

2.6.3 Random cell motility

Images were imported into ImageJ with cell tracking done using the MosaicSuite 2D particle tracking plug-in (Sbalzarini & Koumoutsakos, 2005). XY coordinates for each cell at each time point were exported into MATLAB and a custom-written script used to determine the speed of each cell over time. The average speed of each field of view was then determined with respect to time for comparisons between cell lines.

2.7 Bioinformatics

2.7.1 Identification of actin gene families in other species

Actin genes in commonly used model organisms were identified on Uniprot (http://www.uniprot.org) in conjunction with literature searches. Dictyostelid actin genes were initially identified by using BLAST (https://blast.ncbi.nlm.nih.gov) to search for homology to the actin protein of *Dictyostelium discoideum*. The top hit in this search was then used to re-probe the genome for actin sequences. The 'Identical proteins' feature on NCBI BLAST and/or manual comparison of all potential hits was used to

determine whether genes encode identical amino acid sequences, ensuring that genes were sampled from unique contigs. Genes were identified as 'actin' if annotated as such in the NCBI database or, for less well annotated genomes, those with at least 80% sequence identity to the top hit for actin in that organism.

2.7.2 Calculation of codon adaptation index

Calculation of codon adaption index (CAI) was done using two different software packages. For a small number of sequences the web tool CAIcal was used (Puigbò et al., 2008), while CAI for larger datasets including all protein coding genes was computed using the MBEToolbox in MATLAB (Cai et al., 2005), with the *D. discoideum* codon usage table downloaded from http://www.kazusa.or.jp/. Protein coding gene sequences were downloaded from dictyBase (http://dictybase.org).

2.7.3 Sequence alignment of actin genes

The Bioinformatics Toolbox in MATLAB was used to do a multiple sequence alignment of all 17 genes of the *act8* group in *D. discoideum*. For each gene a 3131 bp sequence was aligned, including the coding sequence (1131 bp) and 1000 bp of up- and downstream flanking sequences. Manual adjustment of highly conserved sequence features was used to more clearly present homologous motifs in the promoter and 3' UTR.

2.7.4 Structural sequence alignment of actin 3' UTRs

Following identification of conserved regions in the 3' UTRs of actin genes, these regions were assessed for the potential to form structural features in mRNA using LocARNA (http://www.bioinf.uni-freiburg.de/Software/LocARNA/) (Will et al., 2007). A short 100 bp region containing both conserved motifs and the intervening sequence was aligned. This tool uses a multiple sequence alignment to look for potential mRNA structures found across all input sequences.

2.7.5 Determination of mapping bias of actin sequences in RNA sequencing experiments

To determine the potential for mapping bias of actin sequence reads during RNA sequencing experiments a custom-written function in MATLAB was used. This function takes the input of a defined sequence window, length x, reflecting the read length in a normal sequencing experiment, and sequentially compares all possible sequences of this length throughout each actin gene to all other *act8* genes (see figure 3.15). 30 bp and 100 bp of upstream and downstream sequence respectively were included to account for sequencing reads derived from UTRs. For each gene, the proportion of 'reads' unique to that gene was calculated, enabling an understanding of the bias which occurs during the mapping process across the gene family.

2.8 Physiological assays

2.8.1 Cellular growth rate

Growth curves were calculated to assess the growth rate of *Dictyostelium* cells. Cells were initially grown in HL5 medium on petri dishes before being transferred to liquid shaking culture and grown to mid-log phase (2×10^6 cells/ml). A new liquid culture was then inoculated from this log phase culture to a starting density of 2-3$\times10^5$ cells/ml. Cells were counted at regular time intervals over the next 72-96 h using a haemocytometer and the mean generation time during log phase growth calculated using

$$MGT = t \times \frac{log2}{log(n_{\text{end}}) - log(n_{\text{beginning}})}$$

where t is the time between observations and n is the number of cells (either at the beginning or end of the time window).

Chapter 3

Bioinformatic analyses of the *Dictyostelium discoideum* actin gene family

3.1 Introduction

Early studies of the actin gene family in *Dictyostelium* sequenced both the protein and coding sequence (Firtel et al., 1979; Vandekerckhove & Weber, 1980; Romans & Firtel, 1985a) and the flanking regions of numerous actin genes (Firtel et al., 1979; McKeown & Firtel, 1981a,b; Kimmel & Firtel, 1983; Romans & Firtel, 1985b). Experimental determination of specific regulatory motifs in two family members (Cohen et al., 1986; Nellen et al., 1986; Hori & Firtel, 1994) highlighted the importance of understanding the contribution of *cis*-regulatory elements to regulatory control of the gene family. Since then, the genome sequence of *D. discoideum* has been determined (Eichinger et al., 2005) which has given us a more detailed view of the whole gene family. A subsequent phylogenetic analysis elucidated the evolutionary history of the gene family and made comparisons to another *Dictyostelium* species (Joseph et al., 2008). In this chapter, building on the important work of the 1980s, I have expanded our understanding of the regulatory control of actin genes in *D. discoideum* with the aid of more recently developed bioinformatic tools.

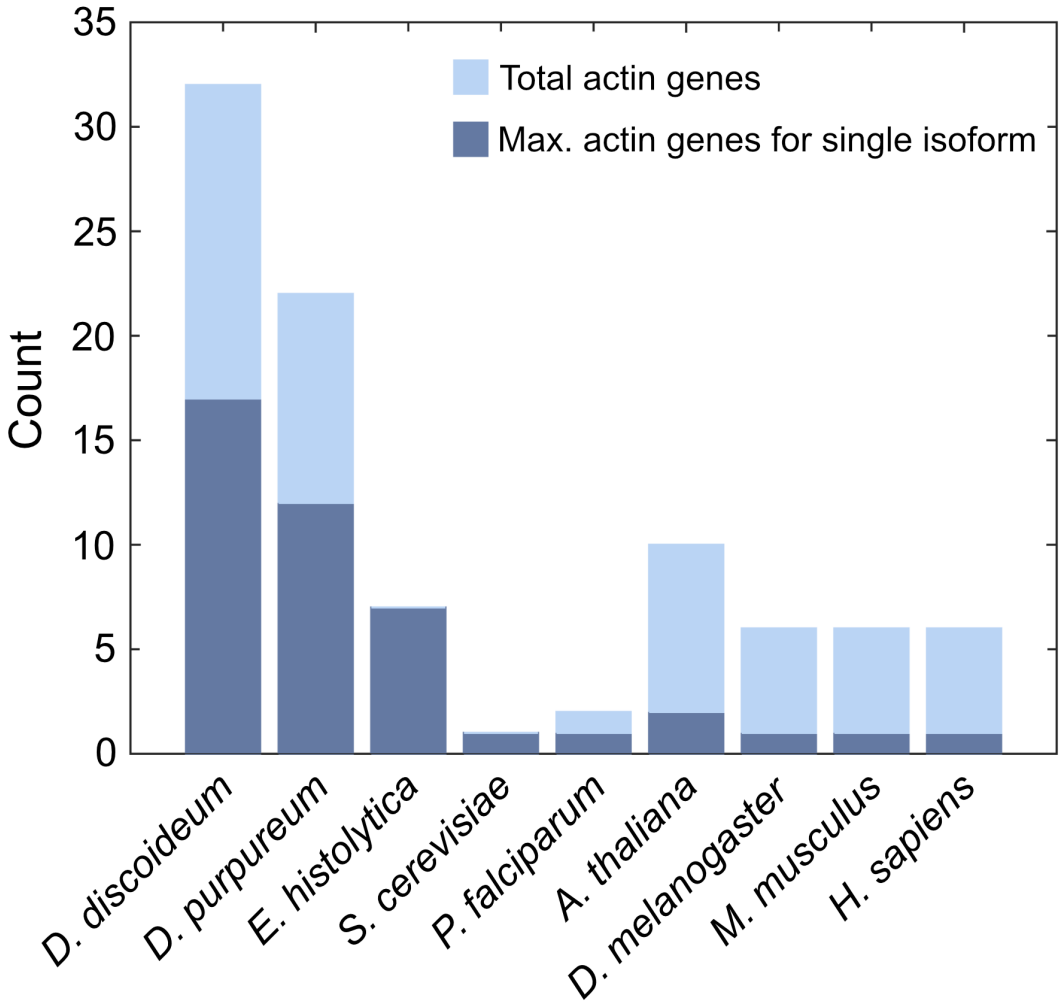

Figure 3.1: Actin gene families in commonly studied eukaryotes. The uniprot database (http://www.uniprot.org) was used to identify actin proteins and the genes encoding them in different model organisms. Total number of genes encoding an actin protein and the maximum number of genes encoding a unique amino acid sequence are shown.

3.2 Actin gene families in other organisms

The actin gene family or 'actinome' of *Dictyostelium discoideum* has been described by Joseph et al. (2008). There are 41 actin or actin-related proteins, and 17 of these actin genes encoding identical protein isoforms (referred to as the *act8* group). To assess how this family compares with other eukaryotes I surveyed the literature and online databases for actin gene families in commonly used model organisms.

As figure 3.1 shows, compared to other eukaryotes *Dictyostelium dis-*

coideum, *Dictyostelium purpureum* and *Entamoeba histolytica* have large actin gene families which contain many genes encoding identical protein isoforms (Joseph et al., 2008; Gunning et al., 2015). In contrast, yeast (*Saccharomyces cerevisiae*, 1), *Plasmodium falciparum* (2), *Arabidopsis thaliana* (10), flies (*Drosophila melanogaster*, 6), mice (*Mus musculus*, 6) and humans (*Homo sapiens*, 6) all have smaller gene families, and only *A. thaliana* has more than one gene encoding a particular actin isoform.

These data could suggest that this phenomenon of larger gene families encoding the same protein isoforms in multiple genes is unique to Amoebozoa species. To address this I explored whether other dictyostelid species have similar actin gene family organisation. Using the *D. discoideum* major actin isoform protein sequence as a probe I used BLAST (https://blast.ncbi.nlm.nih.gov/) to search other dictyostelid genomes available at NCBI for actin genes. I then either used the 'Identical proteins' feature on NCBI BLAST or manually compared all potential hits to determine whether genes encode identical amino acid sequences, ensuring that genes were sampled from unique contigs. Genes were identified as 'actin' if annotated as such in the NCBI database or, for less well annotated genomes, those with at least 80% sequence identity to other actin genes. In fact, some actin genes in *D. discoideum* are only 30% identical to the main isoform (identified as actin proteins by their domain sequence homology in Joseph et al. (2008)) and therefore this could be an underestimate of the size of these gene families. Table 3.1 shows the results of this analysis. While *D. discoideum* has the largest actin gene family of all those surveyed, most dictyostelid species have numerous actin genes, of which several encode identical protein isoforms. This trend appears to hold across all phylogenetic groups of Dictyostelia (Romeralo et al., 2011). From these data it would appear that large actin gene families containing duplicates which encode identical protein isoforms are a general feature of dictyostelid, and perhaps more generally amoeboid, biology.

Species	Phylogenetic group	Genes encoding identical proteins	Total actin genes
Dictyostelium citrinum	4	1	2
Dictyostelium discoideum	4	17	32
Dictyostelium firmibasis	4	3	10
Dictyostelium intermedium	4	7, 2	11
Dictyostelium purpureum	4	12, 3	22
Dictyostelium lacteum	3	3	17
Polysphondlium pallidum	2B	5	9
Acytostelium subglobosum	2A	4	6
Dictyostelium fasciculatum	1	3	13

Table 3.1: Actin genes in Dictyostelia. Genes were identified using BLAST to search homology to *D. discoideum* actin in the currently available genomes of *Dictyostelium* and other dictyostelid species. The total number of actin genes identified with homology greater than 80% are shown, along with the number of genes encoding the same protein isoform, of which there were multiple in *D. intermedium* and *D. purpureum*. The phylogenetic group within Dictyostelia to which these species belong is also shown (Romeralo et al., 2011).

3.3 Genomic organisation of *D. discoideum act8* group genes

While the actinome of *D. discoideum* has been investigated by Joseph et al. (2008) in terms of the diversity across the gene family as well as in relation to other organisms, the *act8* group of genes encoding an identical protein isoform has been less well interrogated. To understand the organisation of this family I first assessed the distribution of these genes in the genome of *D. discoideum*.

A visual representation of the positional information of these genes (Fig. 3.2) highlights some interesting features of the gene family, which are partially and briefly discussed in Joseph et al. (2008). Firstly, the *act8* group genes are distributed across the genome on four of the six chromosomes. Looking more closely, actin genes are organised in clusters on both chromosome 2 and chromosome 5, with clusters of 2, 3 and 4 genes arranged in a head-to-head fashion, as opposed to in tandem (Fig. 3.2). Cluster sizes range from about 10 kb (*act6*, *act19-21*) to 80 kb (*act9*, *act13-14*).

Analysis of duplicate genes in humans has shown that recent duplicates are more likely to be situated close together on the same chromosome, with subsequent rearrangement to other chromosomes taking place relatively slowly (Lan & Pritchard, 2016). Similar phenomena were identified in the *Dictyostelium discoideum* genome (Eichinger et al., 2005). This would suggest that at least some of the duplication events in this system occurred long ago in the evolution of the organism. In addition, this would also suggest that clustered genes are more recent duplicates which is in keeping with phylogenetic analysis of *act8* group coding sequences by Joseph et al. (2008). Overall this implies that *Dictyostelium discoideum* has maintained a large, expanding family of genes encoding identical proteins over long periods of evolution. This would suggest a strong selective pressure both for the particular amino acid sequence (Nowak et al., 1997), but also potentially for the presence of numerous genes, as the long-term maintenance of many identical genes is presumably costly for the cell (Adler et al., 2014) at least in the initial period following a duplication.

3.4 Variable codon adaptation in *Dictyostelium* actin genes

While each of the 17 *act8* group genes encode an identical protein of 376 amino acids in length, the degeneracy of the genetic code means that this can be realised in 2.3×10^{174} different ways. Traditionally, single nucleotide substitutions which direct the usage of synonymous codons have been viewed as 'silent', as they do not change the amino acid sequence of the protein. However, more and more evidence has accumulated to suggest that protein codon usage is both selected for and functionally important (Chaney & Clark, 2015). The nucleotide sequences of the *act8* group genes range in divergence from 2 (*act2* and *act12*, 0.5%) to 108 (*act1* and *act8*, 28.7%) of 377 codons. To test whether these differences have led to significant, and therefore potentially functional, divergence in codon usage within

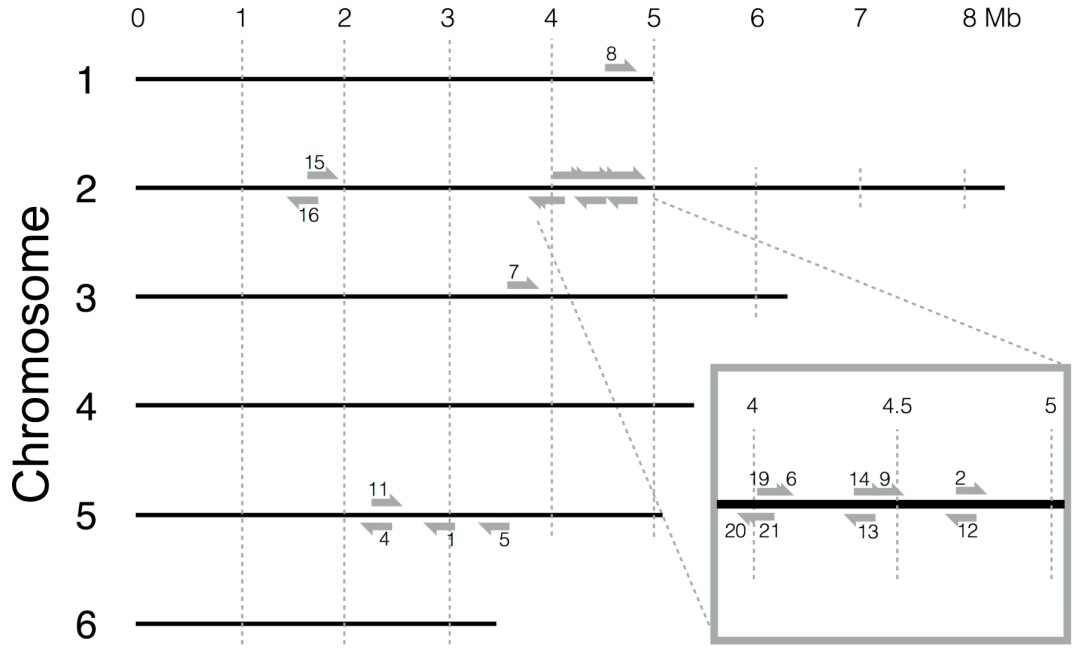

Figure 3.2: Genomic distribution of the *Dictyostelium discoideum act8* group genes. The location and orientation of the *act8* group genes along the length of each chromosome are shown. '1' represents *act1*, '2' is *act2*, and so on. Overlapping labels are arranged as such simply to show the proximity of genes given the scale of the map; the coding sequences do not overlap in reality.

the gene family I calculated the codon adaptation index (CAI) (Sharp & Li, 1987) of all actin genes using CAIcal (Puigbò et al., 2008). This is a measure of how 'adapted' a coding sequence is relative to the genome-wide codon usage bias. If a gene has a CAI of 1 then all the codons used are 'optimal' (the most widely used codon for each amino acid in the genome) and if the value is closer to 0 then most codons are 'non-optimal' and more infrequently used genome-wide (Sharp & Li, 1987).

Comparing actin CAI with RNAseq read counts from vegetatively growing cells (Rosengarten et al., 2015) we find that genes with a higher CAI are more weakly expressed, while lower CAI genes are generally more highly expressed (Fig. 3.3 A). Actin genes appear to be clustered according to this classification rather than exhibiting a strict linear relationship between the two variables (Fig. 3.3 B). One caveat regarding the use of RNA sequencing data here is that the similarity of these actin genes presents a potential

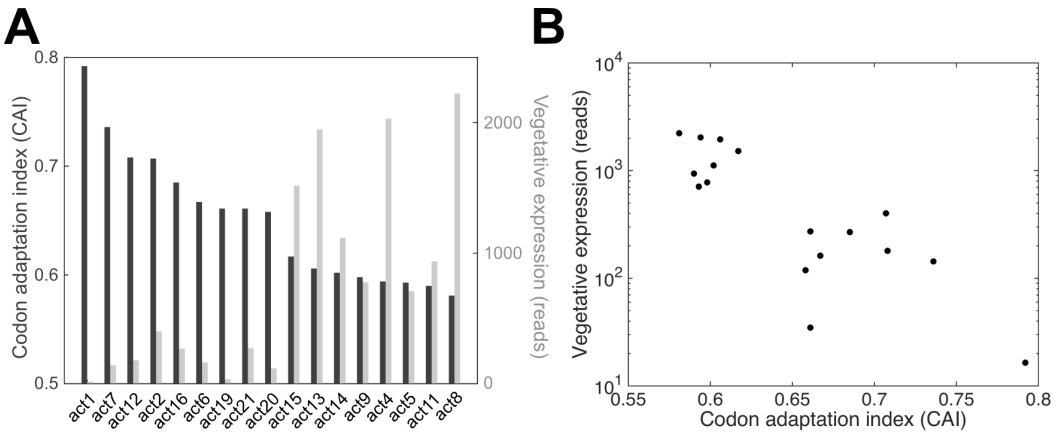

Figure 3.3: Codon adaptation of the actin gene family relative to gene expression. (A) and (B) show the same information of the codon adaptation index (CAI) relative to the expression of each gene in vegetative cells (data from Rosengarten et al. (2015)) as a bar graph (A) or scatter plot (B). CAI was calculated using CAIcal (Puigbò et al., 2008)

problem in the mapping of sequencing reads to the genome, which could affect these results. This is discussed in more detail at the end of this chapter (see section 3.8). The trend shown in figure 3.3 is the opposite of what might be expected, as highly expressed genes have long been thought to include more optimal codons (Sharp & Li, 1987; Jansen et al., 2003). However, more recent evidence has suggested that CAI does not always predict expression levels (Kudla et al., 2009). Indeed, early studies of CAI in *Dictyostelium* using a small sample of highly expressed genes were at odds over the existence of this effect (Warrick & Spudich, 1988; Sharp & Devine, 1989).

To determine whether other highly expressed genes are highly adapted in *Dictyostelium* I calculated the CAI for 12,321 *D. discoideum* coding sequences using the MBEToolbox in MATLAB (Cai et al., 2005). As figure 3.4 shows, highly expressed genes are indeed more likely to be highly adapted to the codon usage of the entire genome. *act8* group genes are also highly adapted compared to the genome average. However, compared to the top 2000 most highly expressed genes, the actins are relatively less well adapted to genome-wide codon usage (Fig. 3.4). Futhermore, compar-

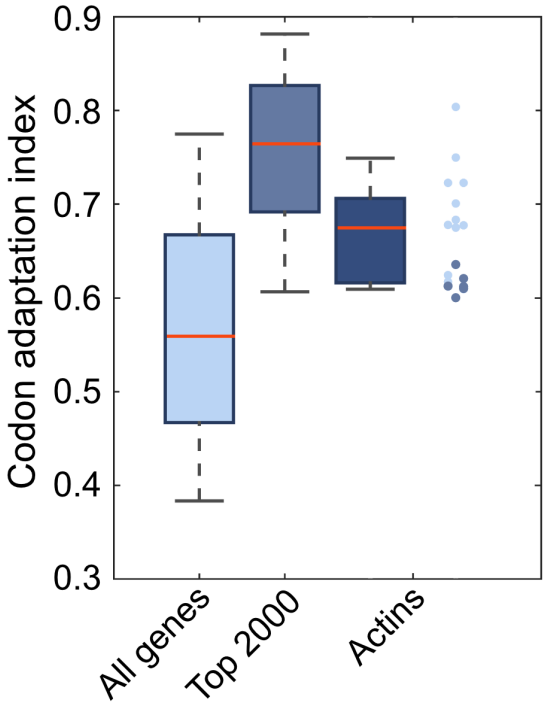

Figure 3.4: Codon adaptation index of all genes in *D. discoideum*. CAI was calculated for 12,321 *D. discoideum* coding sequences downloaded from dictyBase (http://dictybase.org). CAI for the top 2000 most highly expressed genes (from 0 h dataset of single cell RNA sequencing experiments, provided by Vlatka Antolović) and 17 *act8* group genes are also presented. Box plots tails indicate data between 9-91% of the total distribution. Individual actin gene CAI values in dark blue are those found in the 'top 2000' subset.

ing only those actin genes found within the highly expressed subset (dark blue circles in figure 3.4) it is clear that the CAI of highly expressed actin genes is very low compared to the majority of this group of genes. Given that weakly expressed actin genes have a higher CAI, and therefore it is not that actin genes in general have a lower CAI than average, this could suggest that such sequence adaptation has been selected for within the gene family. Higher CAI, and therefore more 'optimal' codon usage, has been demonstrated to increase protein translation rate (Sørensen et al., 1989; Yu et al., 2015; Buhr et al., 2016). Non-optimal codon usage in highly expressed actin genes could be required to slow down translation, potentially to prevent misfolding of these highly expressed proteins (Kim et al., 2015).

3.5 Conservation of common codons in actin protein sequence

Having looked at average differences in codon usage from one gene to another, I next investigated codon usage along the length of the different coding sequences. Firstly, I looked at which codons, if any, are conserved in all 17 *act8* group genes to identify any sites under particularly strong selective pressure. Figure 3.5 A (top panel) shows a sliding average of codon conservation throughout actin genes. Codons at N- and C-terminal regions of the protein, as well as a central region are more strongly conserved in all genes compared to the rest of the protein. This conservation perhaps suggests that these regions are somehow important for protein function.

Which codons are those being conserved? To highlight biases in conserved codons I looked at the codon usage across individual genes. Figure 3.5 A (lower panels) shows sliding averages for *act1* and *act8* of the relative adaptiveness of each codon, known as w. For each amino acid, the most commonly used codon has a w value of 1. The adaptiveness, w, of the remaining codons is defined as the ratio of each codon's usage relative to that of the most commonly used codon. If CUU is the most commonly used codon for leucine at 40% of all leucine positions then its adaptiveness is 1. If CUG codes for leucine in 20% of cases genome-wide, then it has a w value of 0.5. This measure forms the basis of the CAI, which is simply the geometric mean of w (Sharp & Li, 1987). Both *act1* and *act8* follow a similar pattern along the length of the genes; adaptiveness is higher at both termini as well as the central region of the genes. This strongly resembles the profile of overall codon conservation. Indeed, most actin genes show a moderate to strong correlation between the usage of adapted codons and conservation across the gene family (Fig. 3.5 B; $p < 0.01$ for all genes). Therefore, on average more common, or optimal, codons are conserved at these regions of the protein. This, in turn, implies that translational accuracy (i.e. incorporation of the correct amino acid) is important here as rare

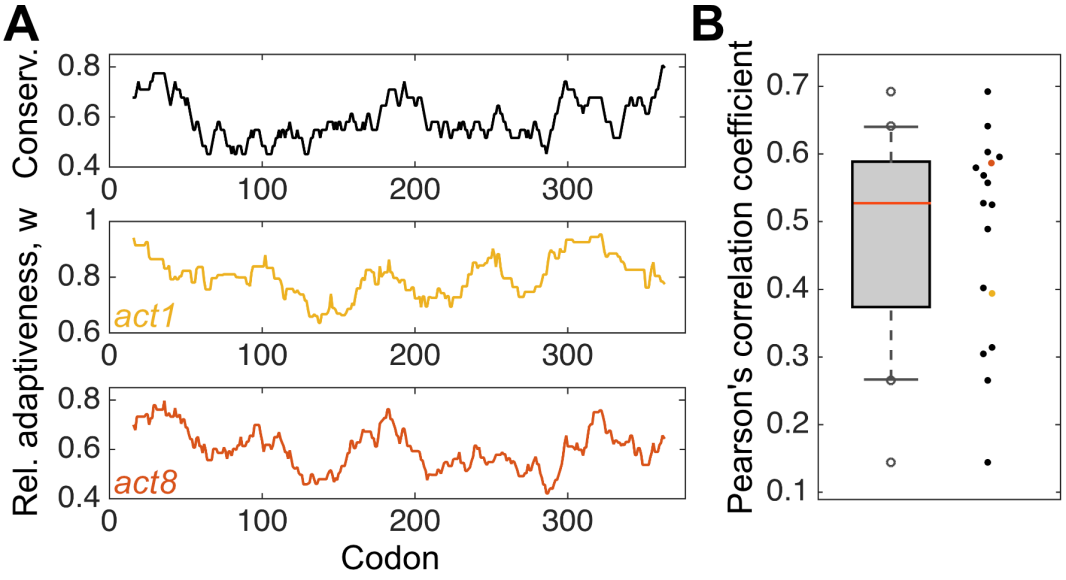

Figure 3.5: Conservation of common codons in actin genes. (A) Conservation across all 17 genes was scored as true (1) or false (0) at each codon position. The top panel shows the sliding average (window = 30) of these scores across the length of the protein. The lower panels show a sliding average (window = 30) of the adaptiveness (or usage) of each codon, w, relative to the usage of the most common codon for that amino acid (as defined in Sharp & Li (1987)) for *act1* (middle panel) and *act8* (lower panel). A larger value for w means the codon at that position is more commonly used throughout the genome. (B) Adaptiveness traces were correlated with codon conservation for each of the *act8* group of genes and Pearson's coefficients plotted as both a boxplot and individual data points. *act1* and *act8* correlation coefficients are shown in yellow and orange respectively.

codons have been shown to cause more 'misread' errors during decoding (Kane, 1995; Kramer & Farabaugh, 2007).

To further investigate the importance of conservation of adapted codons at certain regions of the protein, I looked at potentially relevant structural features for which this translational accuracy might be important. Figure 3.6 A shows a cartoon map of the structure of the ATP-binding pocket of *Dictyostelium* actin (Vorobiev et al., 2003). Highlighted residues are those outlined as important for binding ATP, either directly or via the magnesium cation or water molecules (Vorobiev et al., 2003). Mapping the position of these residues onto the averaged codon conservation profile we see that these structurally important residues are located at the N- and C-termini and

the centre of the linear amino acid sequence (Fig. 3.6 B). This positioning coincides with increased codon conservation and use of common codons (Fig. 3.5). Although this is a weak association, this could suggest that conservation of common codons across the gene family in specific regions of the actin protein may be required to ensure translational accuracy of amino acids involved in forming the ATP binding pocket. The association of common, or optimal, codons with structurally important features, as opposed to intrinsically disordered domains, has previously been demonstrated in other systems (Zhou et al., 2009, 2015).

3.6 Conserved sequence elements in regulatory regions of the gene family

The *act8* group of genes have seemingly been subject to strong purifying selection on coding sequences during *D. discoideum* evolution. How have the regulatory regions controlling the expression of these genes evolved during this time? Joseph et al. (2008) briefly mentioned that analysis of regions up- and downstream of the coding sequences yielded no more information on the proposed putative order of duplications of the *act8* group of genes (compared to analysis done on coding sequences alone). However, this does not tell us about the architecture of the regulatory regions themselves, only that they have likely co-evolved with their associated coding regions. Early studies of actin gene regulatory regions hinted at potential regulatory motifs (Romans & Firtel, 1985b; Nellen et al., 1986; Cohen et al., 1986; Hori & Firtel, 1994) in the promoters of different actin genes. With the sequencing of the *D. discoideum* genome (Eichinger et al., 2005) we now have the ability to examine the flanking regions of all the actin genes in detail.

To investigate the regulatory regions of the actin gene family I performed a multiple sequence alignment of 3131 bp of sequence, including the coding regions and 1000 bp up- and downstream, for each of the *act8* group of genes using the Bioinformatics Toolbox in MATLAB (Fig 3.7). The align-

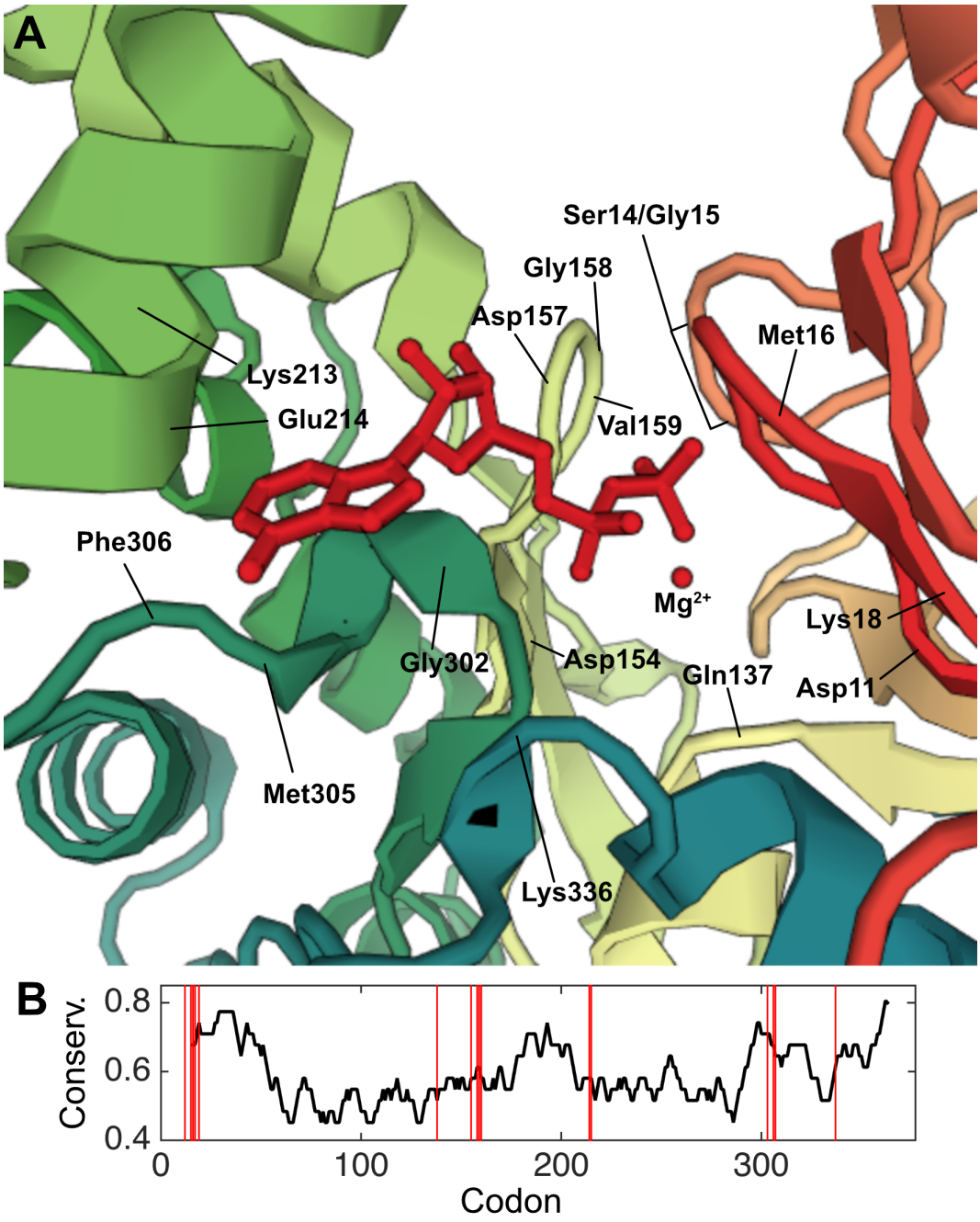

Figure 3.6: Actin ATP-binding amino acids compared to conserved codon regions. (A) A close-up view of the structure of the actin ATP-binding pocket in *Dictyostelium discoideum* (1NLV, RSCB Protein Data Bank;Vorobiev et al. (2003)). The linear sequence of the protein is highlighted by a colour gradient from red (N-terminus) to green (C-terminus). The ATP and Mg_{2+} cation are also coloured red. ATP-interacting amino acids as determined by Vorobiev et al. (2003) are labelled. (B) The position of the ATP-binding residues in (A) are overlaid onto the map of codon conservation from figure 3.5.

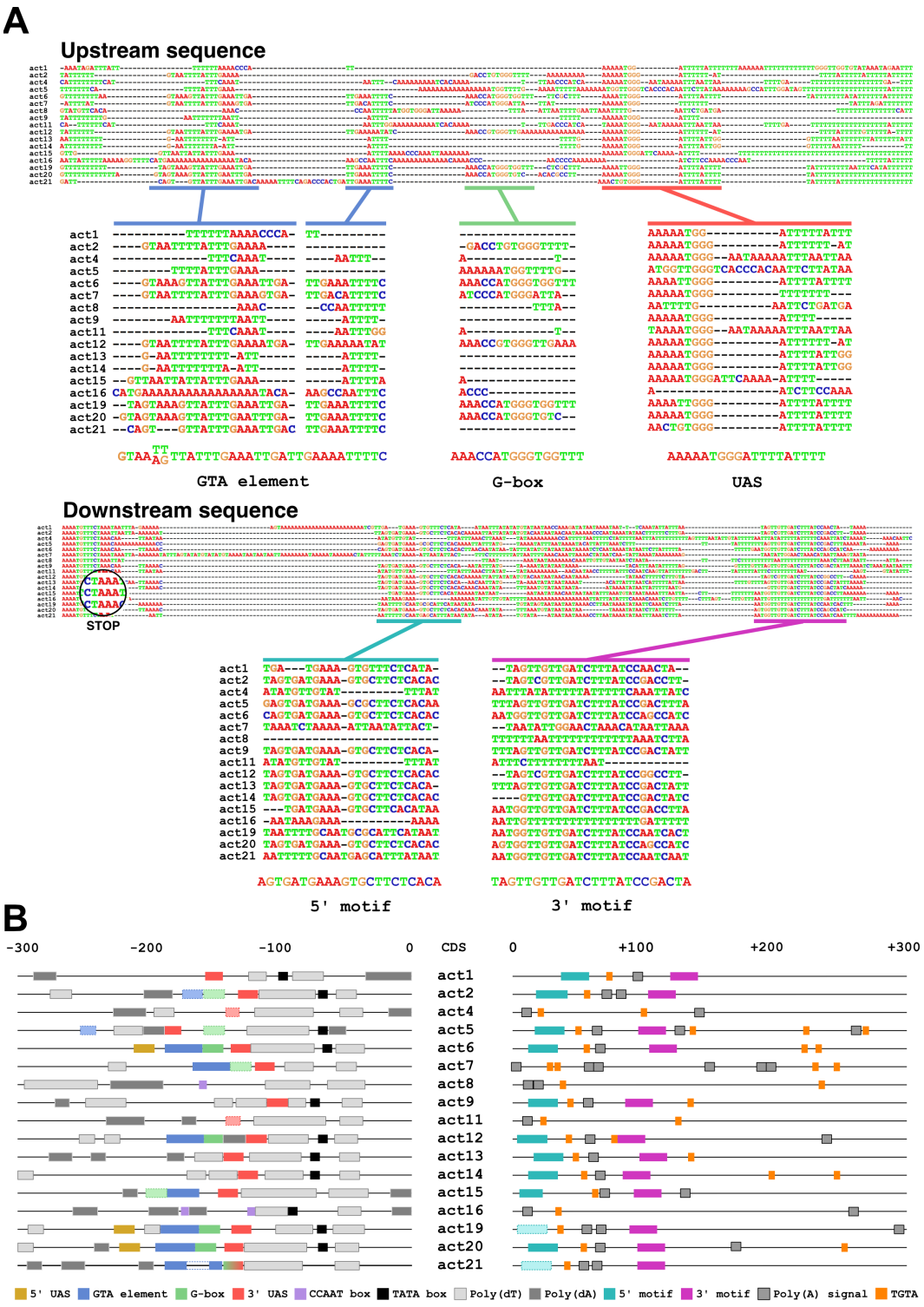

Figure 3.7: Putative regulatory regions of *act8* group genes. (A) Multiple sequence alignment of up- and downstream sequences of actin genes, with manual adjustments in some regions. Close-ups are of regions containing conserved elements of the gene family. (B) Schematic outlining sequence motif positions within actin gene flanking regions based on the alignment in (A). Poly(dA) and poly(dT) tracts are also found in the 3' UTR but are not shown for clarity.

ment was then manually adjusted to improve the arrangement of well conserved regions. As figure 3.7 shows, the alignment revealed the existence of several highly conserved sequence elements in both the promoters and 3' UTRs of these genes. Some of these, including the upstream activating sequence (UAS) (Romans & Firtel, 1985b; Nellen et al., 1986; Cohen et al., 1986; Hori & Firtel, 1994) and the GTA element (Hori & Firtel, 1994, A/T element) have been previously described as being important for gene expression in vegetative growth. From this analysis we can now observe that almost all of the *act8* group of genes have at least one partial UAS, with the exceptions being *act8* and *act16*. A subset of the group (*act2*, *act6*, *act7*, *act12*, *act19*, *act20*, *act21*) also have at least a partial version of the GTA element. It is worth noting here that *act2* and *act12*, and *act6* and *act19-21* are likely recent duplicate pairs (see Fig. 3.2 and Joseph et al. (2008)) which would explain the similarity in promoter sequence.

Beyond compiling a complete picture of the conservation of these previously described elements across the gene family, the alignment analysis also highlighted another sequence motif present in a number of actin genes. A GC-rich palindromic sequence with consensus AAACCATGGGTGGTTT was found in 4 actin genes, with partial or divergent copies in 4-5 more. This was named 'G-box' here because of the distinct similarity with a group of GC-rich elements identified as important for induction in response to cAMP during *Dictyostelium* development (Hjorth et al., 1988; Datta & Firtel, 1988; Pears & Williams, 1988; Pavlovic et al., 1989). These were subsequently identified as targets of the G-box regulatory element binding factor (GBF) (Hjorth et al., 1989, 1990). This G-box element could therefore be involved in regulating the expression of this subset of genes in development. However, the strong induction of actin gene expression during development in *D. discoideum* has been shown to occur in the very early stages (1-3 h) (Tuchman et al., 1974; Alton & Lodish, 1977; Margolskee & Lodish, 1980; Romans et al., 1985), while GBF induces target expression from 6 h of development

onwards (Schnitzler et al., 1994). Given that GBF is now well established as a regulator of mid-late development and is only expressed itself at 4 h of development (Schnitzler et al., 1994, 1995; Brown & Firtel, 1999; Iranfar et al., 2006) it seems unlikely that this putative G-box is bound by GBF in order to activate gene expression.

What is striking about both the UAS and G-box is that the consensus sequences are perfectly palindromic (Fig 3.7 A). This phenomenon of 'dyad symmetry' is found in cis-regulation of gene expression elsewhere, with notable examples including hormone response elements (HRE) in both mammals (Luisi et al., 1991; Khorasanizadeh & Rastinejad, 2001) and plants (Boer et al., 2014), and motifs bound by Rel homology domain (RHD) containing proteins such as NF-κB (Müller et al., 1995; Ghosh et al., 1995), NFAT (Falvo et al., 2008) and Runx1 (Bowers et al., 2010) which are involved in a wide range of processes including the immune response and development. In each case the palindromic motifs, with inverted repeats of 3-6 bp flanking a linker sequence of 1-4 bp, are bound by homodimeric proteins in the major groove of the DNA strand. The architecture of the UAS and G-box sequences found in *Dictyostelium* actin gene promoters closely matches the examples described and therefore we can speculate that these may similarly be targets for homodimeric transcription factors.

Positioned downstream of these putative cis-regulatory elements towards the start of the coding sequences, exists a region containing long tracts of thymine residues, across all actin genes (Fig 3.7 B). This region stretches from the very 3' edge of the UAS to within 30-40 bp of coding sequences which is typically the location of the transcription start site (TSS). Within this there is a gap between two runs of poly(dT) where the sequence is more complex, and is the site of the TATA box for those genes which contain one (the canonical TATAAA(A/T) is specified here). Individual tracts of poly(dT) extend up to 57 bp (*act15*) with little or no interruption from other bases (sequences are marked as poly(dT) in figure 3.7 B only if fewer than

10% non-thymine residues are found within a single stretch). This enrichment at the 5' end of coding sequences is common to most genes in *D. discoideum* (Chang et al., 2012). Poly(dT) tracts were found to be enriched within the linker and at the borders of nucleosome occupancy positions, potentially identifying a functional role for these sequence features (Chang et al., 2012). In keeping with this, the same study found that on average *D. discoideum* genes have a nucleosome depleted region (NDR) centred 54 bp upstream of the TSS, exactly where the TATA box and poly(dT) tracts are found within actin promoters (Fig. 3.7). Poly(dA:dT) tracts have been shown in many other organisms to be less favourable for nucleosome incorporation (Struhl, 1985; Segal & Widom, 2009) and that a 'boundary' such as this in terms of nucleosome formation is likely to reduce the tendency for nucleosome incorporation in flanking sequences also (Kornberg & Stryer, 1988). Therefore, it seems likely that the conserved elements found immediately upstream of the poly(dT) tracts are also more likely to be nucleosome-free. It is worth noting here that while all the actin genes have a region containing stretches of poly(dT), the length of this region, and of the individual tracts, varies from gene to gene. This could enable differences to arise in the regulation of these genes given that poly(dA:dT) tract length is negatively correlated with nucleosome incorporation (Kunkel & Martinson, 1981; Field et al., 2008), which in turn is associated with high levels of gene expression with reduced noise (Kornberg & Stryer, 1988; Sharon et al., 2014). Similarly, the presence or absence of a TATA box has the potential to facilitate differential regulation of actin genes by tuning transcriptional noise as demonstrated elsewhere (Sanchez et al., 2013). This could be particularly important for this system as TATA boxes are enriched in actin genes (12/17 = 70%) compared to the genome-wide average in *D. discoideum* (10-25%) (Chang et al., 2012, G. S. Chang, personal communication).

Overall, using the sequence alignment of the promoter region of all *act8* group genes we find sequence features common to all genes such as a

poly(dT) tract region which is likely involved in proper nucleosome positioning. However, there is also considerable variability in promoter architecture in terms of poly(dT) tract length, TATA box presence, and presence and number of putative transcription factor binding sites which could enable differential regulation of the gene family. This could be important not only in growing cells but also during the developmental life cycle of the organism.

At the 3' end of the genes we see similar strong conservation of potential regulatory elements, accompanied by variability in UTR architecture (Fig. 3.7 B). The multiple sequence alignment shows two highly conserved motifs, each about 20 nucleotides long and situated around 20 and 100 nucleotides downstream of the stop codon respectively. While some of the conserved promoter elements have been described before, neither of these motifs was reported in the early *D. discoideum* actin literature (despite being evident in retrospect, see McKeown & Firtel (1981b); Romans & Firtel (1985b)) and are only mentioned in passing more recently (Muramoto et al., 2012). Neither of these motifs show any significant similarity to the β-actin 'zipcode' found in birds and mammals which is important for polarised localisation and translational control of actin transcripts (Kislauskis, 1994; Condeelis & Singer, 2005; Hüttelmaier et al., 2005). It is conspicuous that one motif always coincides with the other, and that the distance between the two seems relatively constant at around 60 bp. The location of the canonical poly(A) signal sequence AATAAA also appears to be strictly defined. If the 3' UTR of the gene contains the two conserved elements then the proximal poly(A) signal is always found in the intervening sequence. If the 3' UTR does not contain the elements then the first poly(A) signal is found immediately downstream of the stop codon. It is worth noting that 4 of the 6 genes without a TATA box are also lacking the conserved 3' UTR elements. Some UTRs contain several poly(A) signals which could be utilised in alternative polyadenylation (Elkon et al., 2013), a process which has been suggested to occur at almost 70% of human genes (Derti et al., 2012).

Could there be a functional role for these conserved motifs and, if so, what could it be? One possibility is that these could be involved in the processing of the newly transcribed pre-mRNA into mature mRNA. 3' end processing in mammals involves various proteins complexes binding to sequence motifs to direct two reactions of endonucleolytic cleavage and polyadenylation (Shi & Manley, 2015). While the protein components appear largely conserved in yeast and plants, the sequence specificity of the processing sites appears to be different (Millevoi & Vagner, 2010). 3' end processing components and mechanisms have not been experimentally determined for *Dictyostelium* but all genes appear to utilise the canonical mammalian poly(A) signal (Rivero, 2002) and putative homologues for key mammalian proteins such as CPSF can be found in the Uniprot database (http://uniprot.org). Assuming that *D. discoideum* uses similar mechanisms as mammals, the 3' motif downstream of the AATAAA in most actin genes could represent the 'downstream element' (DSE) found in 80% of mammalian transcripts (Shi & Manley, 2015). This sequence is typically immediately downstream of the cleavage site, which is in turn found 10-30 nt downstream of the poly(A) signal (Shi & Manley, 2015), which roughly matches the spacing we see in the actin gene family (Fig. 3.7). While no clear consensus sequence has emerged, a G/U-rich element followed by a U-rich motif are generally thought to enable CstF binding to pre-mRNA, which is required for cleavage but not polyadenylation of the processed transcript (Yao et al., 2012; Mandel et al., 2008). This description, while vague, resembles that of the first 15 bases of the 3' motif shown here, which could explain the strong conservation of this element across the gene family.

Other complexes can bind elements upstream of the poly(A) site to facilitate cleavage. Mammalian cleavage factor I (CFI_m) is thought to bind to UGUA motifs positioned upstream of the poly(A) signal (Brown & Gilmartin, 2003; Hu et al., 2005) which is similar to the efficiency element (EE) in yeast (Mandel et al., 2008). Searching for such a motif in the 3' UTRs of

actin genes in *Dictyostelium* we find a consistent presence of this tetramer around 2-16 nt upstream of the poly(A) signal in genes with the two conserved elements. In the 5 genes without these elements TGTA is found downstream of the proximal poly(A) signal at similar distances. While it might be expected that such a short motif could appear at random in intergenic sequence, two factors appear to suggest otherwise here. Firstly, the fact that between the two conserved motifs the TGTA is found exclusively upstream of the poly(A) signal, in keeping with the proposed role for this motif. Secondly, of the G residues found between the conserved motifs across all 12 genes, 62% are present in a TGTA motif (13/21). In summary, this suggests that a large proportion of the actin gene family in *D. discoideum* contains conserved sequence motifs which are highly similar to mammalian cleavage and polyadenylation elements in their 3' UTRs.

So far I have not addressed a role, if any, for the 5' motif. Structured elements within mRNA molecules are known to influence many different facets of RNA biology (Wachter, 2014) and as structure is often strictly determined by sequence, conserved elements could represent important structural features in actin mRNAs. To test this I used LocARNA (http://www.bioinf.uni-freiburg.de/Software/LocARNA/) (Will et al., 2007), an online tool for simultaneous alignment and consensus structure prediction of multiple RNAs in parallel. The alignment and predicted structural features of the conserved element region for those genes containing both are shown in figure 3.8. Interestingly, both conserved elements are predicted to form hairpin-loop structures, with the base pairs that form the core of the structures strongly conserved across all genes. While RNA structure predictions are notoriously difficult, the strength of this algorithm is that it finds a consensus by comparing all possible structures for each input sequence, which should increase the confidence in the final result. If these structures are found *in vivo* they could be targeted by RNA-binding proteins to regulate localisation, stability or translation of the mRNAs as demonstrated in other systems

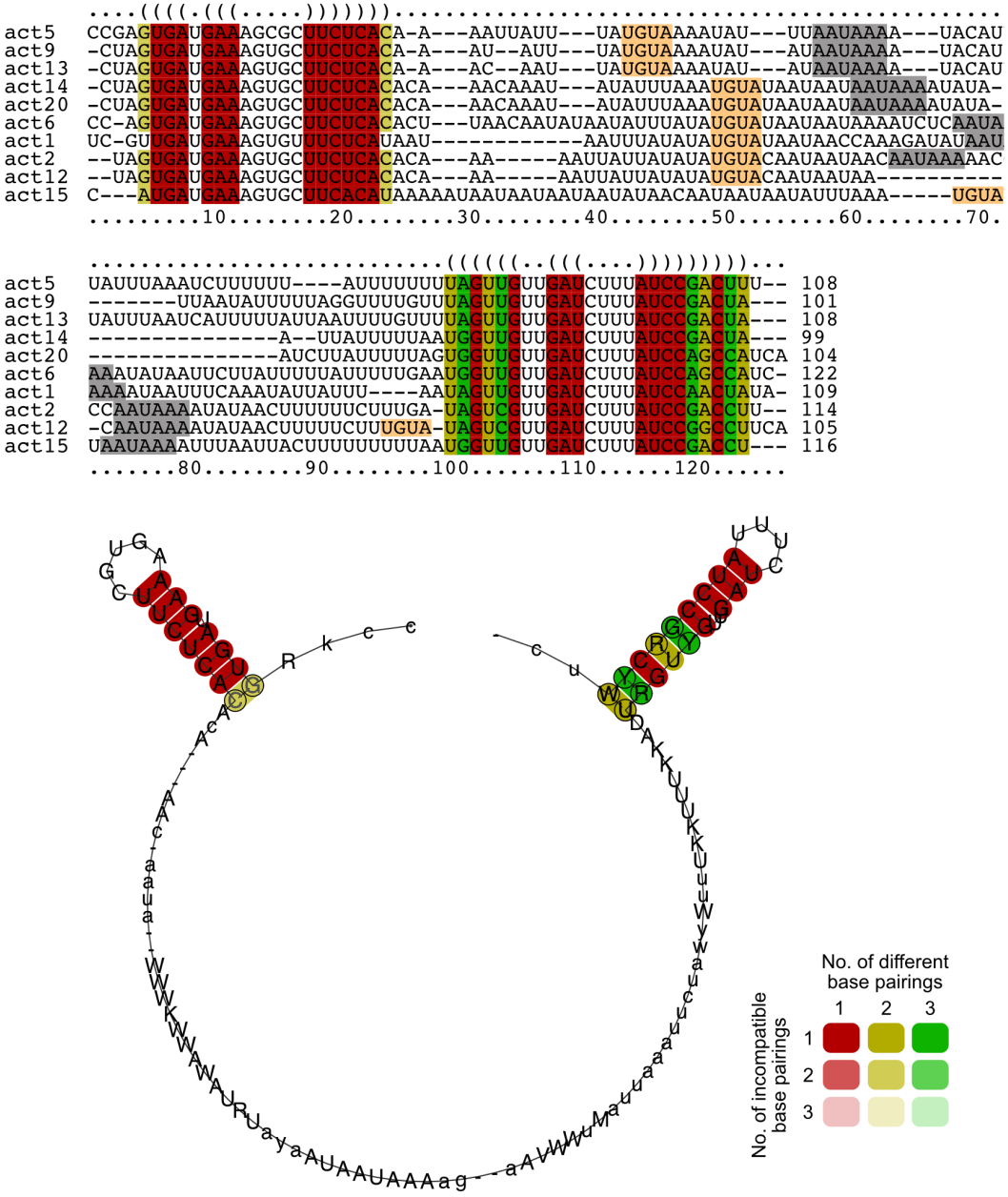

Figure 3.8: Secondary structures in actin gene 3' UTRs. Sequences spanning the region containing the two conserved elements of the actin 3' UTRs were uploaded to LocARNA, a web tool for multiple sequence alignment and structure determination. The output is shown here. Two stem-loop structures were predicted corresponding to the highly conserved motifs. Colours represent number of different base pairings across the input sequences and shades of colour represent the number of incompatible pairs. UGUA and AAUAAA motifs are also shown in orange and grey respectively.

(Svoboda & Cara, 2006). Alternatively, and as the similarity to mammalian DSEs may further imply, the 3' motif could also be used to bring distant cleavage sites into proximity of the 3' end processing machinery located at the poly(A) signal sequence (Wu & Bartel, 2017). In any case, the two classes of 3' UTRs in *Dictyostelium* actin could represent further opportunity for distinct regulation of these genes.

An interesting aside worth mentioning here is that the palindromic elements found at both ends of the coding sequences in several actin genes (Fig. 3.7) could shed light on the duplication mechanism which mediated expansion of the family. Indeed, the tandem inversion duplication (TID) model mentioned in section 1.1.1 (see also figure 1.2) proposes a role for such flanking palindromes in facilitating the duplication of DNA segments during single strand break repair. Furthermore, this mechanism predicts a head-to-head arrangement of the resulting duplicates, which matches the genomic organisation of actin genes seen here (Fig. 3.2). Thus, duplication of *D. discoideum* actin genes may have occurred by TID.

3.7 Actin gene expression during development

3.7.1 Population average measurements of gene expression

Having seen that the promoters of actin gene family members contain a number of different sequence elements, and that these are conserved in varying combinations across the family, I next assessed how these might impact gene expression during the developmental life cycle of *Dictyostelium*. Actin synthesis has been previously shown to be rapidly induced upon starvation and entry into development, before decreasing to below vegetative levels in later development (Tuchman et al., 1974; Alton & Lodish, 1977; Margolskee & Lodish, 1980). To test these findings I analysed *D. discoideum* RNA by Northern blot during development (Fig. 3.9). Relative to ribosomal RNA, actin mRNA levels increase in early development, peaking at 2 hours

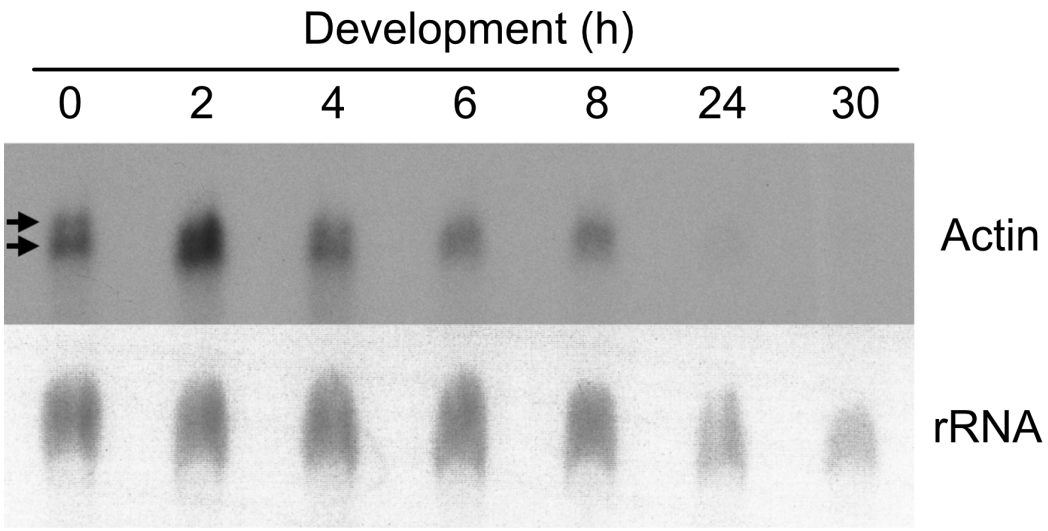

Figure 3.9: Northern blot analysis of actin mRNA during *Dictyostelium* development. RNA from a developmental time course was probed for actin using a 700 bp PCR product generated with primers common to all *act8* group genes. Ribosomal RNA from a gel photo is shown as a loading control. Arrows indicate the presence of two different isoforms of actin mRNA.

before decreasing towards late development. This matches the profile of actin gene expression seen in early studies (Margolskee & Lodish, 1980; Romans et al., 1985). Also evident from this analysis is the presence of two separate bands of actin mRNA (Fig. 3.9, arrows). This has been previously described (Kindle & Firtel, 1978), with the different 3' UTRs outlined in figure 3.7 likely responsible for the different molecular weights of the actin transcripts (Romans & Firtel, 1985b).

While this type of analysis is useful it doesn't tell us whether family members are regulated differently during development. Previous work has attempted to address this question (Romans et al., 1985) but was based on hybridisation of probes to the short 5' UTR sequences of each gene. These regions are more unique than the coding sequences but, as the authors acknowledge, in some cases are almost identical between genes making distinction between them impossible. Modern methods of gene expression analysis, such as RNA sequencing (RNAseq), are much more sensitive and can distinguish highly similar mRNAs based on small regions of unique cod-

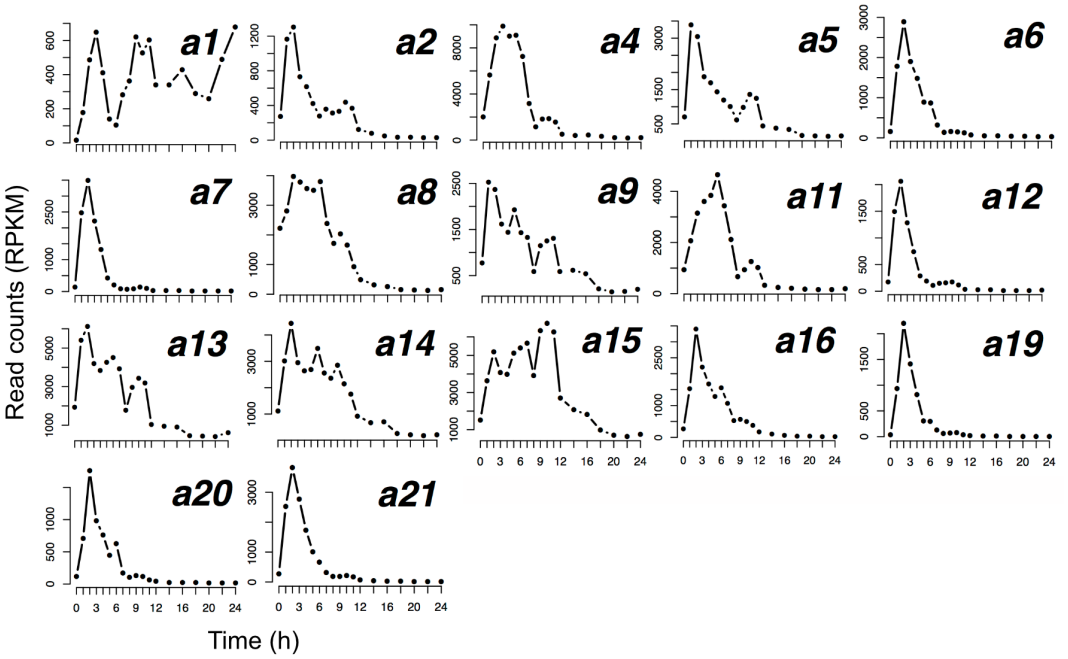

Figure 3.10: RNAseq analysis of actin gene expression during *Dictyostelium* development. Unique sequence reads of *act8* group genes were plotted as a developmental time course (data from Rosengarten et al. (2015)). RPKM is reads per kilobase per million.

ing sequence. Taking advantage of this, I analysed the expression profiles of actin genes during a RNAseq time course of *D. discoideum* development (Rosengarten et al., 2015).

The uniquely mapped read counts of all 17 genes encoding the major actin isoform are shown in figure 3.10. In agreement with the Northern blot analysis, all actin genes are induced upon starvation with a peak of expression between 1-3 hours. Following this, most actin genes show decreased expression during mid development (around 6-10 h) and by 16 h almost all genes show very little expression at all. An exception to this is *act1* which fluctuates during development with peaks during early, middle and late development. However, the relative levels of this transcript are much lower than other actin genes with maximal read counts of around 600 compared to several thousand for other genes (e.g. *act4*, *act8*) and therefore its functional role in the cell may be more limited. Broadly then, actin gene expression during development follows the same pattern, as there are no major

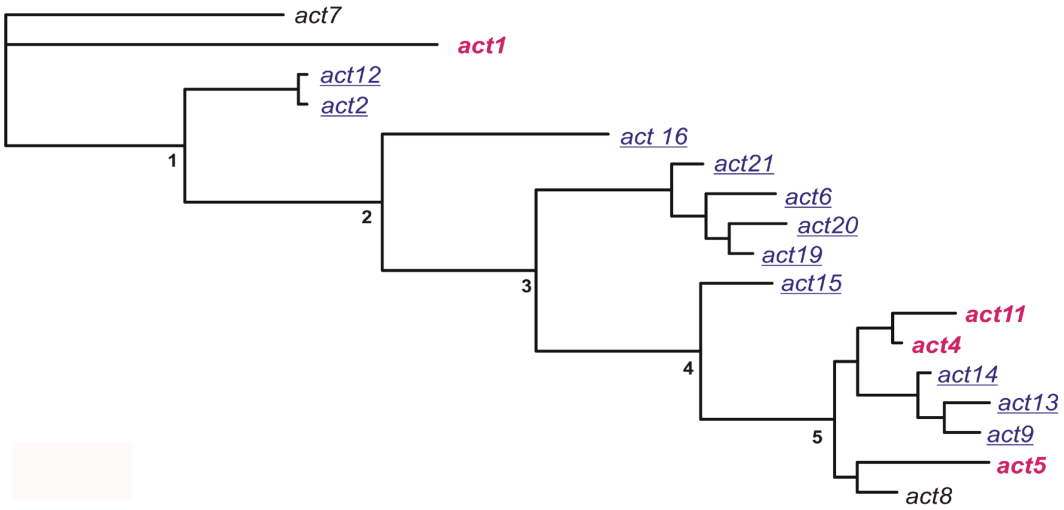

Figure 3.11: Phylogenetic tree of *act8* group genes. The evolutionary history of *act8* group genes (taken from Joseph et al. (2008)). A series of 5 major duplications followed by numerous more recent duplications was predicted to have taken place. Genes in blue and underlined are on chromosome 2, genes in red are on chromosome 5.

divisions across the gene family in terms of timing of expression. However, more subtle differences between family members are revealed upon closer inspection.

Several subgroups of genes can be identified when comparing the different shapes of the developmental expression profiles. For example, *act9*, *act13* and *act14* are all expressed in a series of three descending peaks of expression during the first 12 hours of development whereafter gene activity is minimal (Fig. 3.10). Similarly, *act4* and *act11* have a broad peak, maximal at 3-6 h of development, with a smaller peak at 9-10 h. Finally, *act6*, *act7*, *act12*, *act19*, *act20* and *act21* all have a rapid induction of expression, peaking at 2 h before a similarly precipitous fall in expression to negligible levels at 7 h. Comparing these groupings to predicted recent duplicate clusters by Joseph et al. (2008) (Fig. 3.11) we see that recent duplicates have similar gene expression profiles. Furthermore, the conservation of putative regulatory elements in promoter regions is strongly correlated with developmental expression (Fig. 3.7, 3.10). For example, genes with a full-length GTA element followed by a G-box, such as *act6* (Fig. 3.7), show a sharp peak of

expression in early development (Fig. 3.10). This would suggest that the diversity of promoter elements explored over evolution are responsible for the different developmental expression patterns of actin genes, as opposed to any effects from local chromatin environments. However, whether the subtly different profiles have specific roles in the developmental process is unclear.

3.7.2 Single cell analysis of gene expression

Technological developments of the last five years have enabled the genome-wide profiling of gene expression from single cells within a population using single cell RNA sequencing (scRNAseq) (Kolodziejczyk et al., 2015). To determine whether actin genes are able to compensate for other family members, and therefore whether this might explain the maintenance of such a large gene family, I analysed data from a scRNAseq time course of *Dictyostelium* development (with assistance from Vlatka Antolović and Agnès Miermont) (Antolović et al., 2017).

Cells were developed and harvested at 0, 3, 6 and 14 h of development, with 2 replicates and 70-100 cells in total from each time point (experiments done by Agnès Miermont). Figure 3.12 shows the distribution of gene expression of single cells at each time point for four different genes, two actin genes and two genes which are known to be induced in early development, *carA* and *csaA* (Saxe et al., 1991; Noegel et al., 1986). The scRNAseq time course closely matches the trends seen using other analysis methods, with actin genes peaking at 3 h of development, while carA and csaA increase from 0 through 6 h (Fig. 3.12).

To test whether certain actin genes are able to compensate for others in the same cell I compared the expression of all 17 *act8* genes with each other for each developmental time point (Fig. 3.13). A clustered heat map of all correlations shows that most genes are either moderately or strongly correlated in their expression across the whole family, from a population of single cells for all developmental time points (Fig. 3.13 B). The genes which are more weakly correlated with other actin genes throughout development

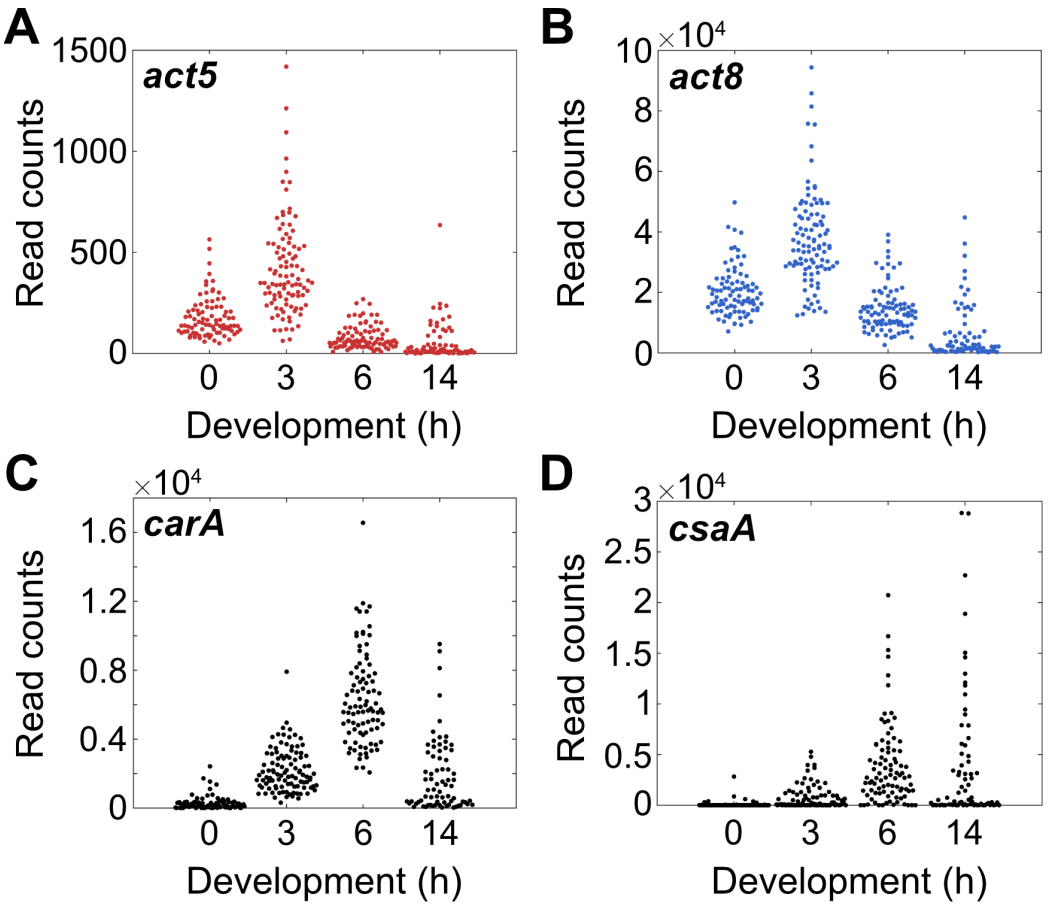

Figure 3.12: Example of scRNAseq reads during development. Read counts from single cell RNA sequencing experiments in *Dictyostelium* development for different genes are shown (Antolović et al., 2017). Marked points represent individual cells. *carA* and *csaA* are developmentally induced genes.

are more distantly related to to the rest of the family (see Fig. 3.11). Given that these genes also appear to be more weakly expressed this could suggest that these genes are subject to less stringent regulation than others family members. Overall, given that on the whole actin gene expression is correlated in single cells it would seem unlikely that these genes are able to buffer the expression of each other. Rather, it seems that individual cells are coordinating actin expression levels as a group, and therefore generating higher and lower-actin cells, as opposed to some optimum intermediate value. This effect could be at least partially due to differences in cell size within the population as it is well known that cells are able to scale transcript

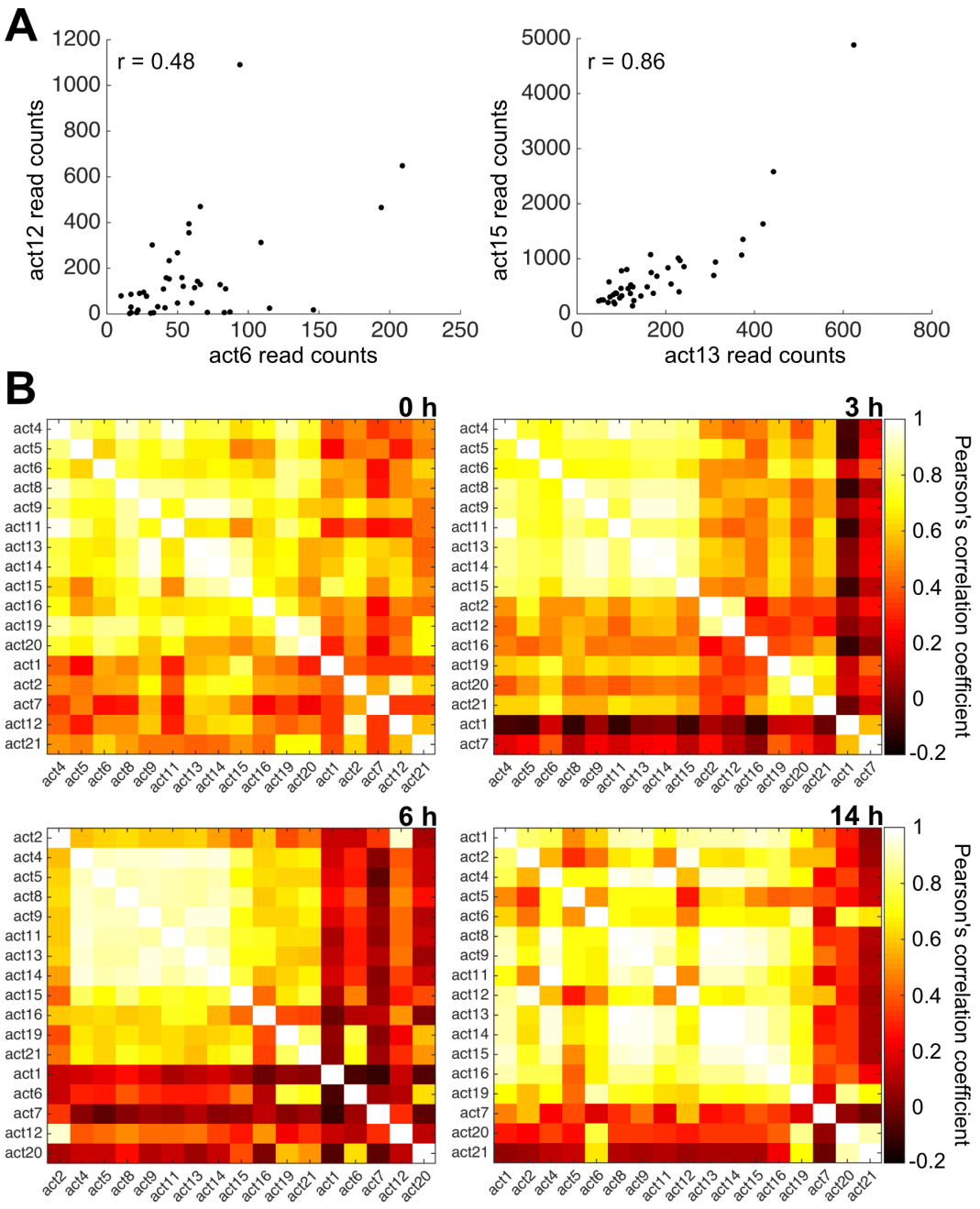

Figure 3.13: Correlations in expression of actin genes within populations of single cells. (A) Examples of correlations of scRNAseq reads from two actin gene pairs at 0 h of development. (B) Heat maps of Pearson correlation coefficients comparing all actin genes with each other at 4 developmental time points. Plots were clustered using k-means clustering.

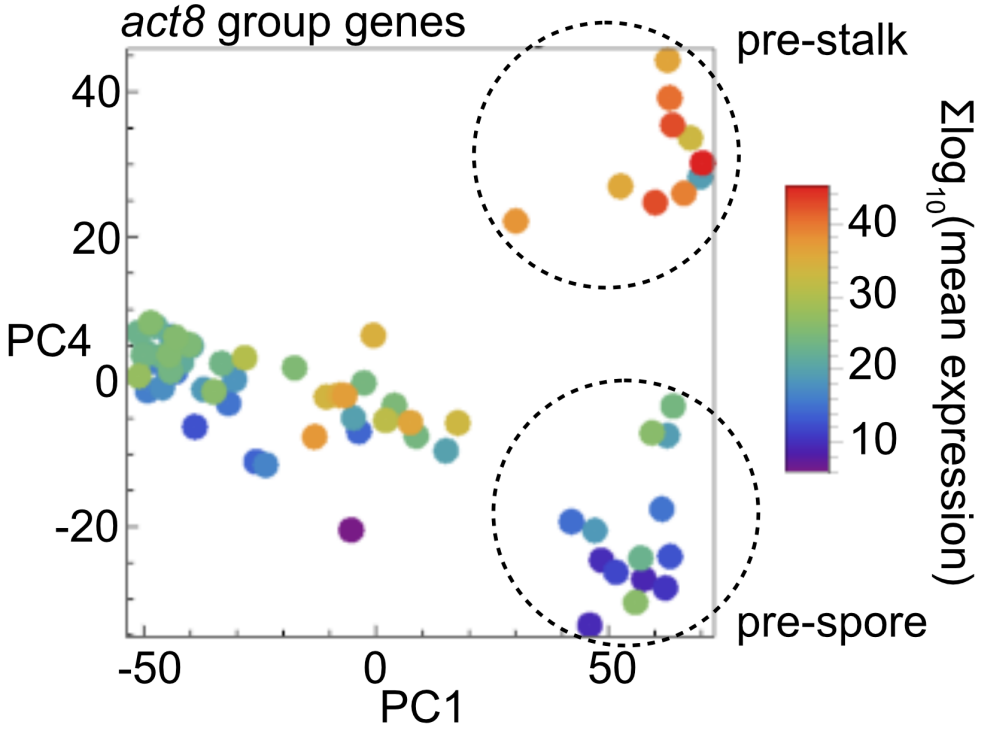

Figure 3.14: Cell-type specific expression of actin genes in late development. Principle component analysis of 14h development scRNAseq dataset. Two subgroups of cells were defined as pre-stalk or pre-spore based on the expression profiles of known markers of these cells types. The mean expression of *act8* group genes is shown for each cell.

numbers according to their volume (Marguerat & Bähler, 2012). However, cell size is unlikely to explain all of the variability in gene expression as volume in *Dictyostelium*, while variable, does not normally range over more than an order of magnitude as is seen with read counts in figure 3.13 A (Bonner & Frascella, 1953; Waddell, 1988). To be conclusive though, these data would need to be normalised by cellular volume.

Could the different expression tendencies within a cell population also be due to cell-type specification during development? Our developmental time course covers early development (0, 3 and 6 h), where cells are globally modulating their gene expression (Van Driessche et al., 2002; Rosengarten et al., 2015; Antolović et al., 2017) but not yet defined by distinct cell fates, and a mid-late developmental time point (14 h) where cells are in tipped mound or slug stages and are strongly expressing markers of differ-

entiation (Van Driessche et al., 2002; Rosengarten et al., 2015). Using a principle component analysis (done by Vlatka Antolović) to investigate the relative contribution of actin gene expression to specific cell types in the 14 h dataset we see that actin genes are strongly associated with pre-stalk cells compared to pre-spore cells (Fig. 3.14). In this analysis, the genes contributing to PC1 were principally developmental genes, while genes with loading values contributing most strongly to PC4 were cell-type specification markers. Similar cell-type specific expression of actin genes in late development has been demonstrated previously (Alton & Brenner, 1979; Coloma & Lodish, 1981; Tsang et al., 1982; Mehdy et al., 1983; Barklis & Lodish, 1985), which further supports the above data. This pattern is consistent for all individual actin genes, and therefore it appears unlikely that different actin genes are specifically upregulated in different cell types, at least at this developmental time point.

3.8 Mapping bias of actin sequencing reads

An important consideration when interpreting the above data is that all RNA sequencing results are quantified in terms of 'read counts' or metrics derived from these counts. Gene-specific read counts are those sequencing reads which can be unambiguously and uniquely mapped to a distinct gene locus. As mentioned previously, in the context of a gene family this can be problematic as a large proportion of the 'multireads' derived from highly similar transcripts will be discarded as they cannot be uniquely mapped to a single locus (Li et al., 2010a).

To test how this might bias the RNAseq results presented above I attempted to calculate the proportion of reads derived from a specific actin gene which could be uniquely mapped. For each actin gene, and for each possible 'read' of a defined length ('x' in figure 3.15) within the transcript I searched for identity across the rest of the gene family. Figure 3.15 shows the proportion of the total number of unique 'reads' which can be mapped to a particular gene. Clearly, the proportion of uniquely mappable reads is

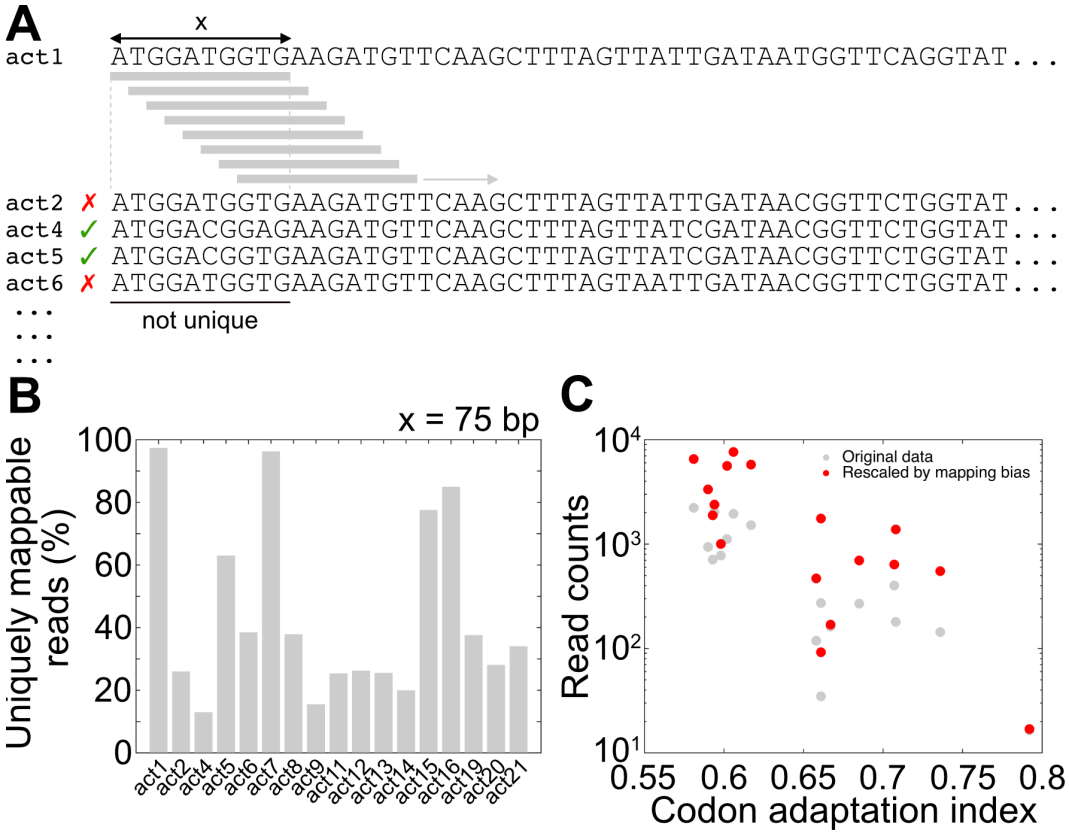

Figure 3.15: Mapping bias for actin genes in RNA sequencing data. (A) Actin genes were assessed for uniqueness within the gene family by searching for exact sequence matches of a defined 'read length', x, across all other actin genes. A sequence was defined as unique only if there were zero exact matches in the rest of the gene family. All possible 'reads' for each gene were assessed including UTRs. (B) The proportion of sequence reads found to be unique is plotted for a window (or read length) of 75 bp. (C) RNAseq data from figure 3.3 is adjusted to account for the mapping bias in (B).

not equal across the gene family. *act1* has a relatively unique sequence, which means that when a read length of 75 bp is used (as in Antolović et al. (2017)) about 97% of *act1* reads can be mapped to this locus. In contrast, for *act4* only about 12% of reads will be uniquely mapped. Actin gene expression levels as measured by RNA sequencing are therefore likely higher than illustrated above. However, the pattern of gene expression through development should be unaffected as the mapping bias should occur regardless of experimental conditions. This basic simulation does not account for use of paired-end sequencing, used in many RNAseq experiments (although not

Antolović et al. (2017)), which would be expected to reduce the number of discarded reads. Replotting the data from figure 3.3 after accounting for this mapping bias (Fig. 3.15 C) show that while the overall trends in the data are unchanged, there are considerable differences in the read counts for some actin genes after this correction. This analysis highlights the need to be cautious when using RNA sequencing to explore gene expression in multi-gene families, and should encourage the use of complementary methods to answer important scientific questions. This forms the basis of the following chapter.

3.9 Summary

In this chapter, using bioinformatic methods, I have explored the actin gene family of *Dictyostelium discoideum* in the context of other eukaryotes, the variability of both coding and flanking sequences of the genes, and how these differences may impact the developmental expression of the gene family as a whole. I have shown that the actin gene family of *D. discoideum* is unusually large, and contains many more genes encoding a single protein isoform, compared to most other eukaryotic model organisms (Fig. 3.1). However, I also showed that this family organisation seems to be conserved in other dictyostelid species (Table 3.1) and therefore may be an important part of *Dictyostelium* biology. I then explored the genomic distribution of the *act8* group of genes and found that separation of genes on different chromosomes suggests a relatively ancient gene family, yet smaller clusters also imply some more recent duplications (Fig. 3.2).

Codon usage of the *act8* group was found to be variable with highly expressed genes using more weakly adapted codons and weakly expressed genes using highly adapted codons (Fig. 3.3), a result which contrasts with the traditional relationship of codon usage and gene expression. This was further supported by analysis of all *D. discoideum* genes which showed that highly expressed actin genes have very low CAI values compared with the most highly expressed genes in vegetative cells. I also showed that con-

servation of codon usage is stronger in certain areas of the protein and that conserved codons are more likely to be common codons (Fig. 3.5).

To investigate the potential for differential regulation of the gene family I generated a global picture of flanking sequence architecture to identify conserved sequence elements (Fig. 3.7). Several putative regulatory elements were identified including two motifs with perfect dyad symmetry, hinting at a role in binding a dimeric transcription factor complex. Some elements such as the UAS were found in almost all genes while other elements including the TATA box were variably distributed, which may permit distinct control mechanisms of actin gene expression. At the 3' end of genes I found two strongly conserved sequence elements (Fig. 3.7) which were predicted to form stem-loop structures (Fig. 3.8). Again, variable conservation of these in different actin genes may permit differential regulation of the genes. In particular, the arrangement of one conserved element and another short 4 bp motif relative to the poly(A) signal shows strong similarity with mammalian signatures of 3' end processing of mRNA (Fig. 3.7).

To understand how these putative regulatory elements might impact gene expression in the context of *Dictyostelium* development I analysed data from both bulk population and single-cell RNA sequencing experiments (Figs. 3.10, 3.12). Developmental expression of individual actin genes is broadly similar but with subtle differences likely explained by variable promoter architecture (Fig. 3.10). Using scRNAseq I showed that it is unlikely that multiple actin genes are required as a buffer for noisy gene expression, as actin levels in *D. discoideum* cells seem to be coordinated globally (Fig. 3.13). I also showed that actin gene expression is more closely associated with stalk cell specification, as opposed to spore, within the developing organism (Fig. 3.14). Finally, I showed that actin gene mapping bias in RNA sequencing experiments is highly likely, meaning caution should be taken with this kind of experimental data and highlighting the need for alternative measures of gene expression in this system (Fig. 3.15).

Chapter 4

Gene expression dynamics of *D. discoideum* actin genes

4.1 Introduction

In the previous chapter RNA sequencing analysis indicated that the general pattern of actin gene expression in development is broadly similar (Fig. 3.10). As a result, differential regulation of actin genes during development appears unlikely to represent a functionally important explanation for gene family expansion. However, gene expression, and more specifically transcription, can be regulated over much shorter time scales than those of developmental transitions. Indeed, work over the last decade has shown that most genes are not subject to strict binary control, being either 'on' or 'off', but are in fact transcribed in bursts or pulses of activity lasting minutes at a time (Golding et al., 2005; Chubb et al., 2006; Raj et al., 2006; Lenstra et al., 2016). Therefore, in this chapter, I assessed whether these short-term dynamics of gene expression are differentially regulated in actin genes.

Importantly, I also showed in the preceding chapter that while use of methods such as RNA sequencing to study the actin gene family in *Dictyostelium* are effective in exploring general trends of gene expression, issues with potential bias in the mapping process make it hard to quantitatively study individual genes. Therefore, it was important here to be able to specifically target particular actin genes for study. Here I have made use

of different reporter systems to precisely measure the activity of individual actin genes at different stages of protein production, thus enabling a detailed look at the different regulatory mechanisms at play within the gene family.

4.2 Generation of cell lines to monitor dynamics of actin gene transcription

To monitor the transcription dynamics of individual actin genes I generated cell lines containing an array of 24 repeats of an MS2 RNA bacteriophage sequence which, when transcribed into mRNA, forms a stem-loop structure which is specifically bound by the MS2 coat protein (MCP) (Romaniuk et al., 1987; Valegård et al., 1994). Expressing a MCP-GFP fusion protein in cells containing the tagged gene enables fluorescent labelling of mRNAs derived from the gene of interest (Bertrand et al., 1998). In live cells, a fluorescent 'spot' is visible at the site of transcription within the nucleus (Fig. 4.1) which can be seen to 'pulse' or 'burst' over time (Golding et al., 2005; Chubb et al., 2006). This bursting behaviour represents the transcriptional activity of the gene (Muramoto et al., 2012; Corrigan et al., 2016).

I generated three constructs to target *act1*, *act6* and *act8* genes with the MS2-containing cassette (Fig. 2.1) while an *act5*-MS2 cell line has also been previously used in the lab (Muramoto et al., 2012; Corrigan et al., 2016). The four actin genes chosen have a range of promoter architectures (Fig. 3.7), expression levels and developmental profiles (Fig. 3.10). The regulation of *act5*, *act6* and *act8* has also been studied to some extent previously (Romans et al., 1985; Nellen et al., 1986; Hori & Firtel, 1994). Upon removal of the blasticidin resistance gene (*bsr*) using the Cre-LoxP system (to ensure transcripts use the natural terminator) and expression of the MCP-GFP construct, I imaged the different cell lines in growth medium to assess the dynamic transcriptional activity of the target genes. As the snapshots in figure 4.2 show, the different actin genes exhibit very different transcriptional activities. In general, taking the intensity of a transcriptional

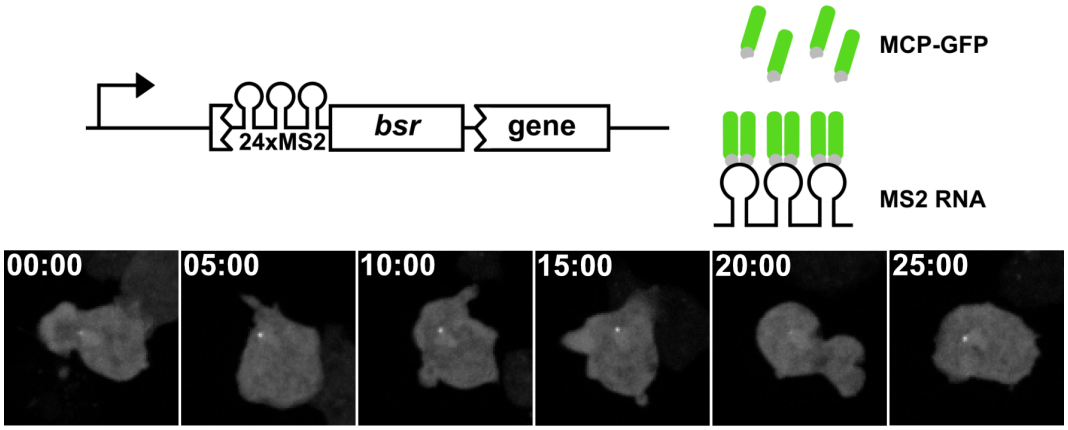

Figure 4.1: Imaging gene activity using the MS2 system. To visualise transcription in living cells, a cassette containing 24x MS2 stem-loop sequences along with a *bsr* resistance gene is inserted into the 5' end of the gene interest (top panel). Upon correct integration, the *bsr* gene is removed by Cre-LoxP recombination. Subsequent expression of a plasmid containing the MS2 coat protein (MCP) fused to GFP, which specifically binds to the stem-loops, enables visualisation of a 'transcription spot' in the nucleus (top and bottom panels). Time values are in minutes. Time-lapse imaging of these cells shows these spots to be dynamic or pulsatile in nature, with signals lasting for minutes to tens of minutes at a time, depending on the gene of interest (bottom panel).

spot to represent the magnitude of gene activity, a visual assessment of the four actin genes matches data from hybridisation and sequencing experiments (Romans et al., 1985, and Figs. 3.10, 3.12). No *act1* transcription spots were detected in any of the live imaging experiments done (Fig. 4.2 A) which suggests that this gene is either inactive or active below the detection threshold of the system (previous work suggests a higher detection threshold for this technique, around 5 mRNAs, than others such as smFISH (Corrigan et al., 2016)). Activity of *act5* appears to be variable within the population, with some cells transcribing strongly while others show no activity at all (Fig. 4.2 B). Very few cells show activation of *act6* transcription, but when a spot does appear in these cells it can be almost as bright as those of *act5* (Fig. 4.2 C). Finally, *act8* appears to be active in almost all cells, with bright transcription spots demonstrating strong gene activity (Fig. 4.2 D). These qualitative differences in activity, with *act1* the weakest, followed by *act6*, *act5* and *act8* being the most active, match those seen in the RNAseq

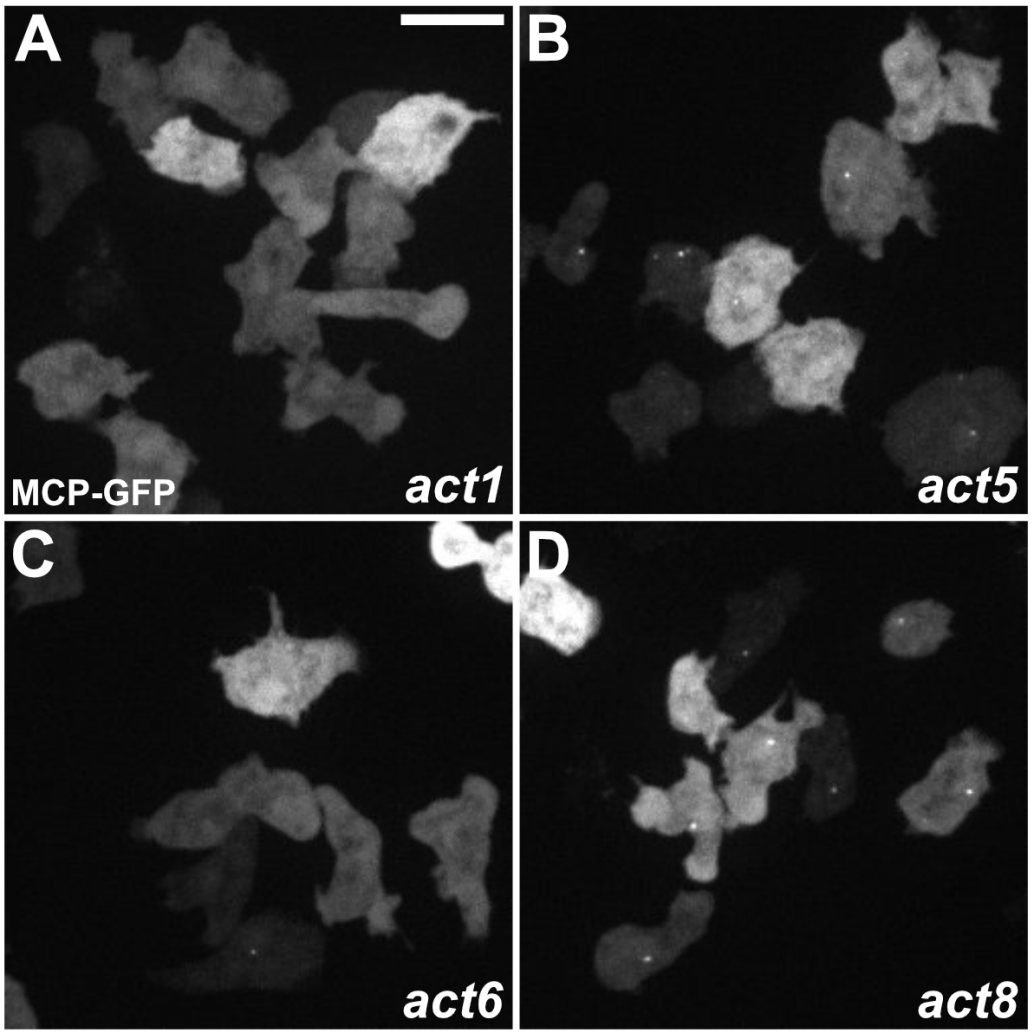

Figure 4.2: Assessment of transcriptional activity of multiple actin genes by MS2 tagging. (A) *act1*, (B) *act5*, (C) *act6*, (D) *act8* genes were targeted in a cell line containing an H2B-mCh knock-in using the method described in figure 4.1 (see section 2.1.8.1). Transcription spots are visible for all genes (except *act1*), with different proportions of cells expressing at any one time. Scale = 12 μm.

data (see 0 h time points in figure 3.10). Therefore, as a measure of actin gene expression this system broadly agrees with other methods. However, the ability to monitor this transcriptional activity over time in live cells also enables us to assess the dynamic behaviour of these genes, which can reveal distinct regulatory features of gene expression (Muramoto et al., 2012).

4.3 Transcriptional dynamics of actin gene expression in vegetative cells

To investigate how actin genes might differ in transcriptional dynamics I imaged MS2-tagged cell lines (not including *act1*) in normal growth medium for 30-90 minutes and analysed movies using a custom-built pipeline in MATLAB written by Adam Corrigan (see section 2.6.1, described in more detail in Corrigan et al. (2016)). As the representative examples of spot intensity traces in figure 4.3 show, actin genes display dynamic changes in spot intensity, and therefore transcriptional activity, over time. However, the different genes display different dynamic behaviours. Both *act5* and *act6* show pulsatile transcription which results in sudden bouts of activity followed by periods of apparent inactivity. This is strongly reminiscent of transcriptional bursting, thought to be the dominant mechanism of gene expression in most eukaryotes (Chubb et al., 2006; Raj et al., 2006; Suter et al., 2011; Dar et al., 2012; Bahar Halpern et al., 2015b). While the spot intensity is greater for *act5* compared to *act6* (Fig. 4.3), qualitatively, both genes appear to use this bursting mechanism, with only the burst size or magnitude differing between the two (one can imagine raising the detection threshold for *act5* and thus seeing a similar transcription spot profile to *act6*). In contrast, while *act8* gene activity fluctuates over time, it is far less variable than the other actin genes with spot intensities more tightly distributed around the mean (Fig. 4.3). This potentially represents the use of two different transcriptional mechanisms driving expression of the same protein.

The image analysis software used in this study enables automatic segmentation, tracking and extraction of spot intensities for entire populations of cells (Corrigan & Chubb, 2014; Corrigan et al., 2016). Visualising actin gene activity of hundreds of cells gives a clearer picture of the regulation of the family within a dynamic, cooperative population. Figure 4.4 shows the spot intensity traces of all tracked cells from 4 different experiments (on 3 different experimental days). The bursting behaviour of *act5* and *act6* are clearly

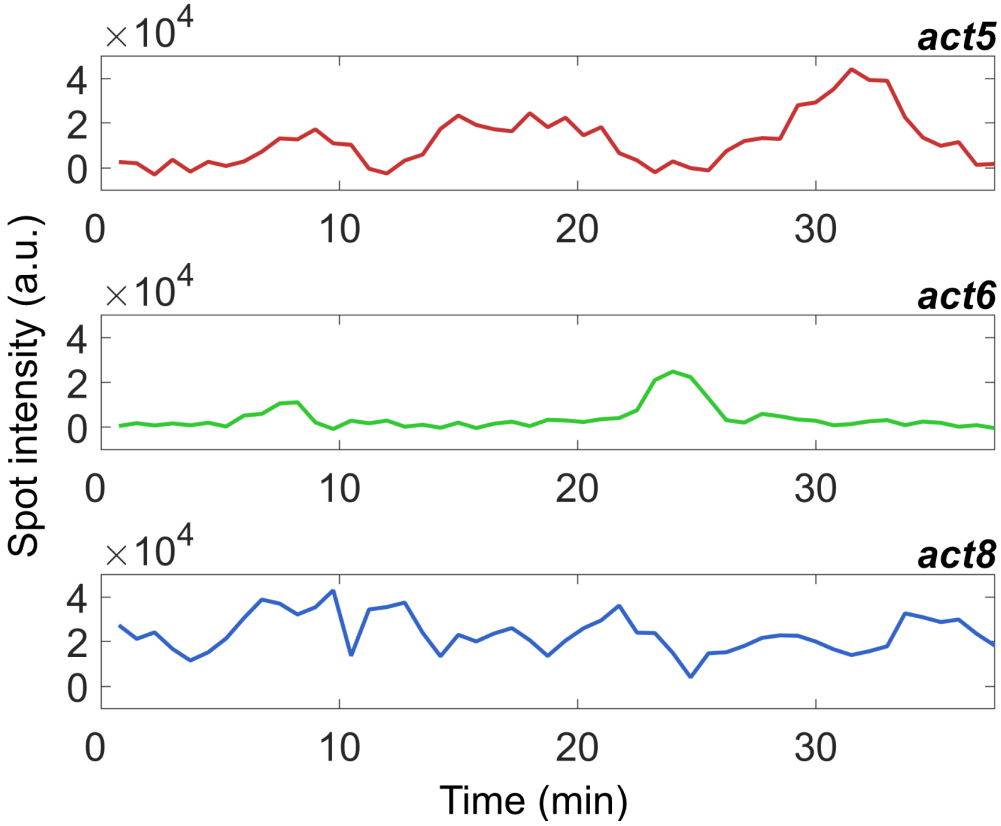

Figure 4.3: Example of actin gene transcription dynamics in individual cells. Spot intensity traces for three individual cells with the MS2 system targeted to different actin genes are shown. Spot intensities are normalised to the nuclear background intensity due to significant variability between cells (see figure 4.2). Movies were captured with a frame interval of 45 s. Custom-written software (Corrigan et al., 2016) automatically segments and tracks cells and then identifies transcription spots, with minimal supervision required.

visible in these plots (Fig. 4.4 A, B). On average *act5* spot intensity is higher than *act6* but both display short periods of strong gene activity interspersed with periods of relative inactivity. Transcriptional activity is not only variable for an individual cell, but also between cells, with some showing no activity whatsoever while others display several transcriptional bursts within a period of 30 minutes. In contrast, *act8* activity is more stable over time, but is also more homogeneous between cells compared to *act5* and *act6* (Fig. 4.4 C). These observations suggest that (at least) two types of transcriptional mechanisms are employed to generate actin mRNA, with *act8* transcribing constitutively (and therefore perhaps generating basal levels of actin), while

other genes such as *act5* and *act6* are more 'bursty', perhaps demonstrating responsiveness to temporally variable stimuli within a dynamic signalling environment. This could be an important explanation for why the gene family has simultaneously expanded while faithfully preserving protein sequence. Previous studies have also described variable bursting dynamics in different genes but these differences were measured on a timescale of hours rather than minutes (Suter et al., 2011), or compared the kinetics of very different groups of genes (Muramoto et al., 2012).

To quantitatively characterise the different transcriptional behaviours of actin genes I firstly plotted the distribution of spot intensities for all tracked frames as a probability density function (PDF) for each gene. Figure 4.5 shows a representative example from one experiment. Given that transcription spots are not detected in *act1*-MS2 cells, the *act1* distribution illustrates the noise in spot intensity measurement, which derives from the fact that the analysis software attempts to identify a transcription spot in all cells, regardless of transcriptional activity. In fact, referring to *act1* as a kind of negative control for gene activity is quite useful, as it highlights the proportion of cells in which the other actin genes are likely to be inactive, or active below the detection threshold, at any one time (Fig. 4.5).

The spot intensity distributions quantitatively describe the qualitative assessment of dynamic behaviours made from figure 4.4. The transcriptional bursting of *act5* and *act6* are clearly visible in the long tails of the distributions, with *act5* able to reach much higher spot intensities than *act6* (and therefore presumably higher initiation rates, see Corrigan et al. (2016)). The *act8* gene activity PDF is much closer to a normal distribution with a mean spot intensity of around 2×10^4 (Fig. 4.5). The differences between the distributions are captured nicely by measurements of variability in the system. *act5* and *act6* are significantly more variable than *act8*, as shown by both the squared coefficient of variation, $\left(\frac{\sigma}{\mu}\right)^2$ and Fano factor, $\frac{\sigma^2}{\mu}$ in figure 4.6 (paired Student's t-test). These measures of noise are normalised by the

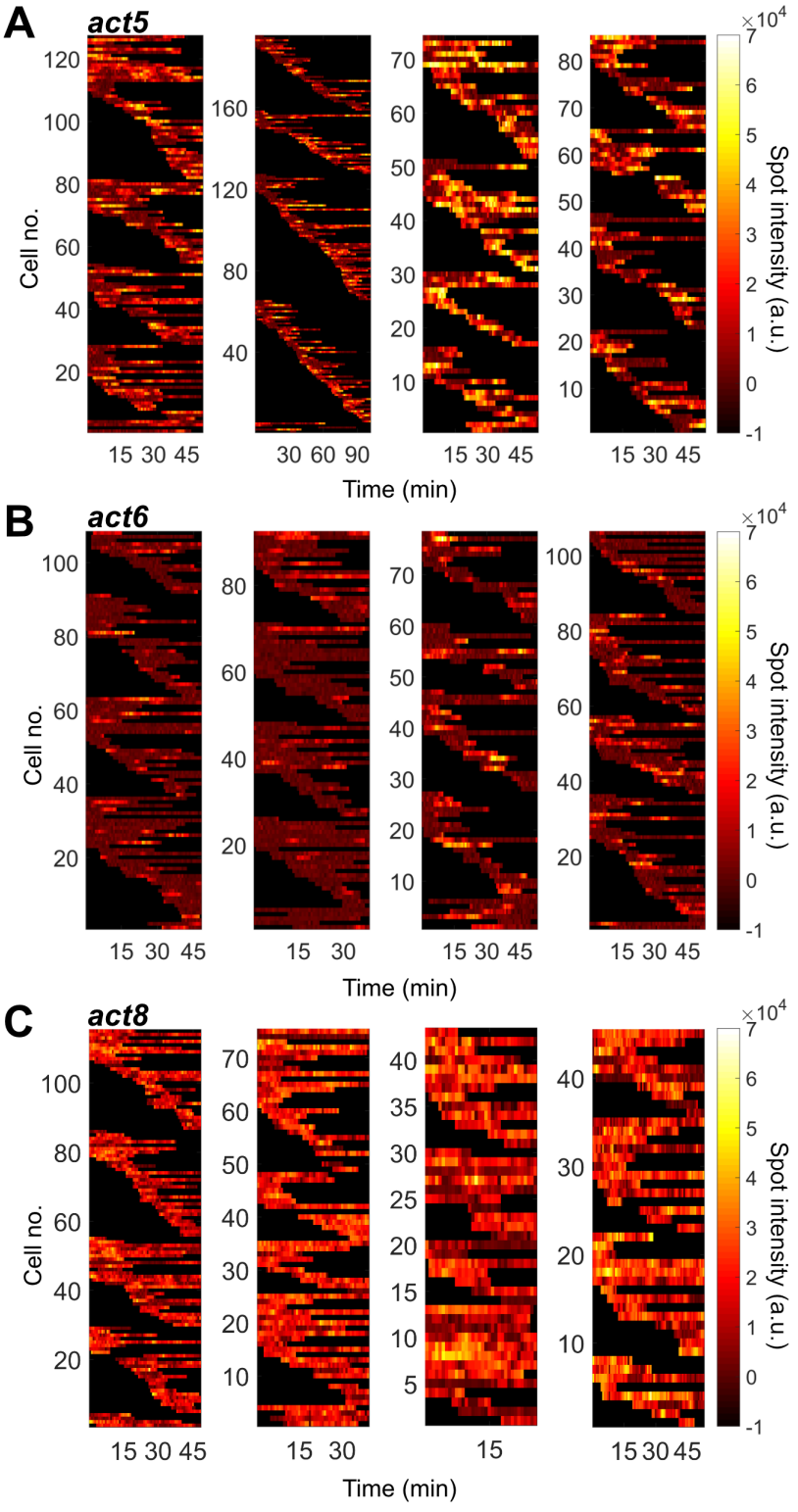

Figure 4.4: Actin gene transcription dynamics. For each actin gene - (A) *act5*, (B) *act6* and (C) *act8* - all transcription dynamics data are shown comprising four movies captured over three experimental days. Each heat map represents one movie, with coloured areas showing tracked cells from four fields of view. Each row represents the spot intensity trace from a single tracked cell, as in figure 4.3. Yellow and white data points represent strong transcriptional activity whereas dark reds indicate little to no activity. Black regions represent time points at which the tracked cell has either left or not yet entered the field of view.

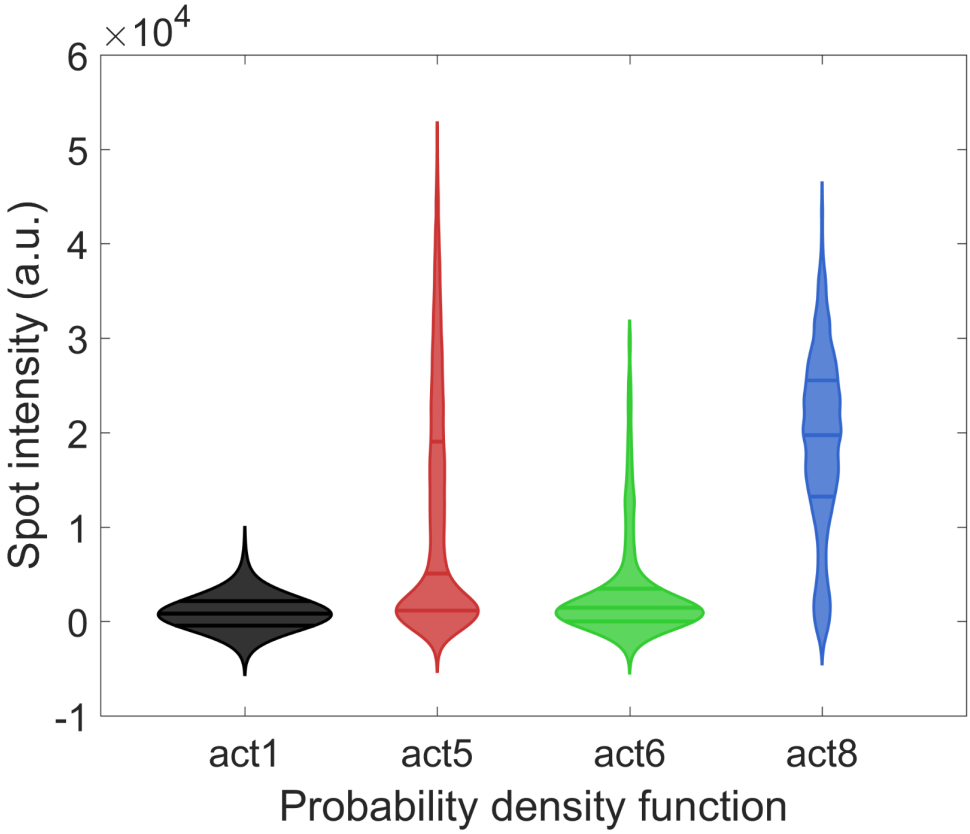

Figure 4.5: Spot intensity distributions of *Dictyostelium* actin genes. For each actin gene a smoothed probability density function of all spot intensity data points was generated. Horizontal lines indicate median and upper and lower quartiles. Negative values arise from the background correction, as in the absence of a clear spot the detection software will mark a random position in the nucleus which, when normalised, may become negative in value. Representative distributions from one experiment are shown.

mean in order to take account of the fact that variance scales with mean expression (Bar-Even et al., 2006; Sharon et al., 2014).

To understand the mechanisms by which the different actin genes are controlled I separated each movie shown in figure 4.4 into individual fields of view (FOV) and calculated the average spot intensity (across all cells) for each FOV. The variability in average spot intensity levels across different FOV enables us to query how changes in gene activity are brought about mechanistically. Measuring the noise (C_v^2) and noise strength (Fano factor) for a gene relative to changes in expression can indicate whether burst

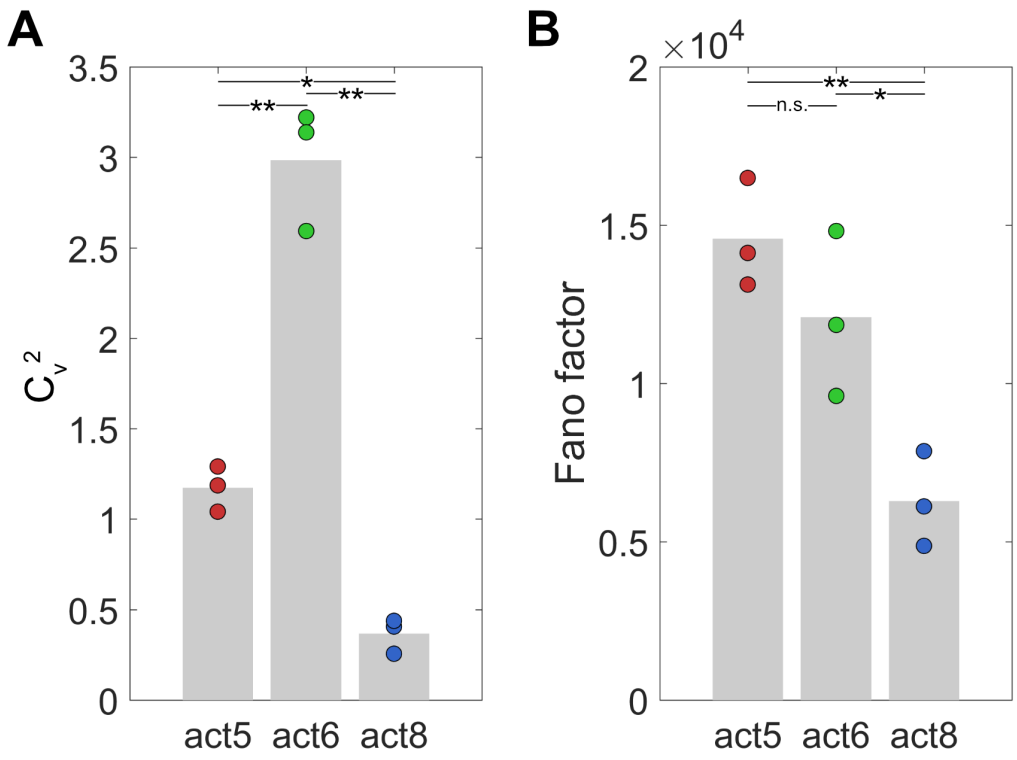

Figure 4.6: Variability of actin gene transcription dynamics. (A) Squared coefficient of variation $\left(\frac{\sigma}{\mu}\right)^2$ and (B) Fano factor ($\frac{\sigma^2}{\mu}$) were calculated for all spot intensity data from three experimental days. Mean values are grey bars, while individual data points are coloured circles. Paired Student's t-test was used to test for significance, although Welch's (unpaired) unequal variance t-test was also used and gave the same result unless indicated in the text. N.s. = p >0.05, * = p <0.05, ** = p <0.01.

frequency or size is preferentially modulated to bring about such change (Carey et al., 2013). Plotting these parameters against average spot intensity for the three actin genes reveals some clear trends (Fig 4.7). All three genes show a strong negative correlation between average spot intensity and noise (Spearman's rank correlation, *act5*: r = -0.92, p = 0, *act6*: r = -0.79, p = 0.0005, *act8*: r = -0.83, p = 0.0001) (Fig. 4.7 A). This would suggest that all three actin genes alter their activities by modulating transcriptional burst frequency (Carey et al., 2013). In contrast, only *act6* spot intensity is significantly correlated with noise strength (Spearman's rank correlation, *act5*: r = -0.13, p = 0.62; *act6*: r = 0.83, p = 0.0001; *act8*: r = -0.36, p = 0.17) meaning *act6* potentially modulates both burst frequency and size

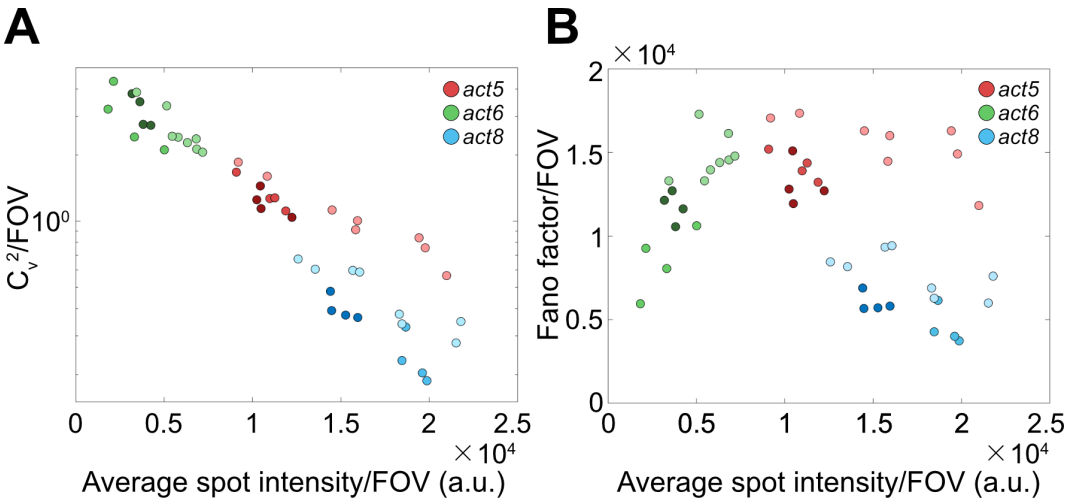

Figure 4.7: Noise-mean relationship of actin gene transcription dynamics. (A) Correlation between squared coefficient of variation, or noise, and averaged spot intensity per field of view for actin genes. (B) Correlation between Fano factor, or noise strength, and averaged spot intensity per field of view for actin genes. Relationships between these different variables supposedly reflect the ability to modulate particular bursting parameters (Carey et al., 2013). Shades of each colour represent different experimental days.

to effect changes in gene expression, while *act5* and *act8* modulate burst frequency alone.

It is also important to note here that despite global noise being known to decrease with gene expression levels (Elowitz et al., 2002; Bar-Even et al., 2006; Hornung et al., 2012; Sharon et al., 2014) the difference in noise between *act5* and *act8* seen in figure 4.6 is unlikely to be simply due to this difference in expression level. Figure 4.7 A shows that for a given average spot intensity value, *act5* noise is consistently higher than *act8*. Conversely, the noise-mean relationship of *act5* and *act6* can probably be explained by the same exponential function, given that the data appear to lie on the same linear regression line (Fig. 4.7 A). Therefore, the difference in C_v^2 between *act5* and *act6* in figure 4.6 A is probably explained by this phenomenon of reduced noise at higher expression levels.

Overall, these data show that actin genes in *Dictyostelium* display at least two different types of transcription dynamics, with highly variable,

strong bursting behaviours of *act5* and *act6* contrasted by a less variable, more constitutive activity of *act8*. The data presented above suggests that *act5* and *act6*, and possibly *act8* although this is not clear from these results alone, are controlled by transcriptional bursts and therefore the contrasting dynamic behaviours are likely a consequence of differential regulation of bursting parameters. Further analysis suggests that while all genes modulate burst frequency to alter gene activity, *act6* can also modulate burst size to increase its activity.

4.4 Determining the effects of promoter sequence on transcriptional dynamics

What are the determinants of the different dynamic behaviours of *D. discoideum* actin genes? Many factors acting over a wide range of length scales are known to control gene expression. Cis-regulatory elements at the promoter (Sharon et al., 2012, 2014), chromatin modifications (Suganuma & Workman, 2011), genomic context (Michalak, 2008) and nuclear organisation (Fraser & Bickmore, 2007) have all been implicated in the regulation of gene expression, to different degress. However, the individual contributions of each of these to transcription dynamics are unclear due to extensive crosstalk between these putative controlling influences (Hershberg et al., 2005; Batada et al., 2007). Here I specifically investigated the role of promoter sequences in regulating bursting dynamics. To do this I generated cell lines where the promoters of the two actin genes with the most distinct transcriptional bursting patterns, *act6* and *act8*, have been exchanged at the endogenous loci (Fig. 2.3). These genes are situated on different chromosomes (Fig. 3.2). As shown in figure 4.8, the transcription dynamics of these modified cell lines were markedly altered from those of the endogenous genes. The presence of the *act8* promoter at the *act6* gene (‘A8P-A6G’ in figure 4.8 and hereafter) increases the spot intensity, and reduces the intra- and inter-cell variability of the population, akin to that seen

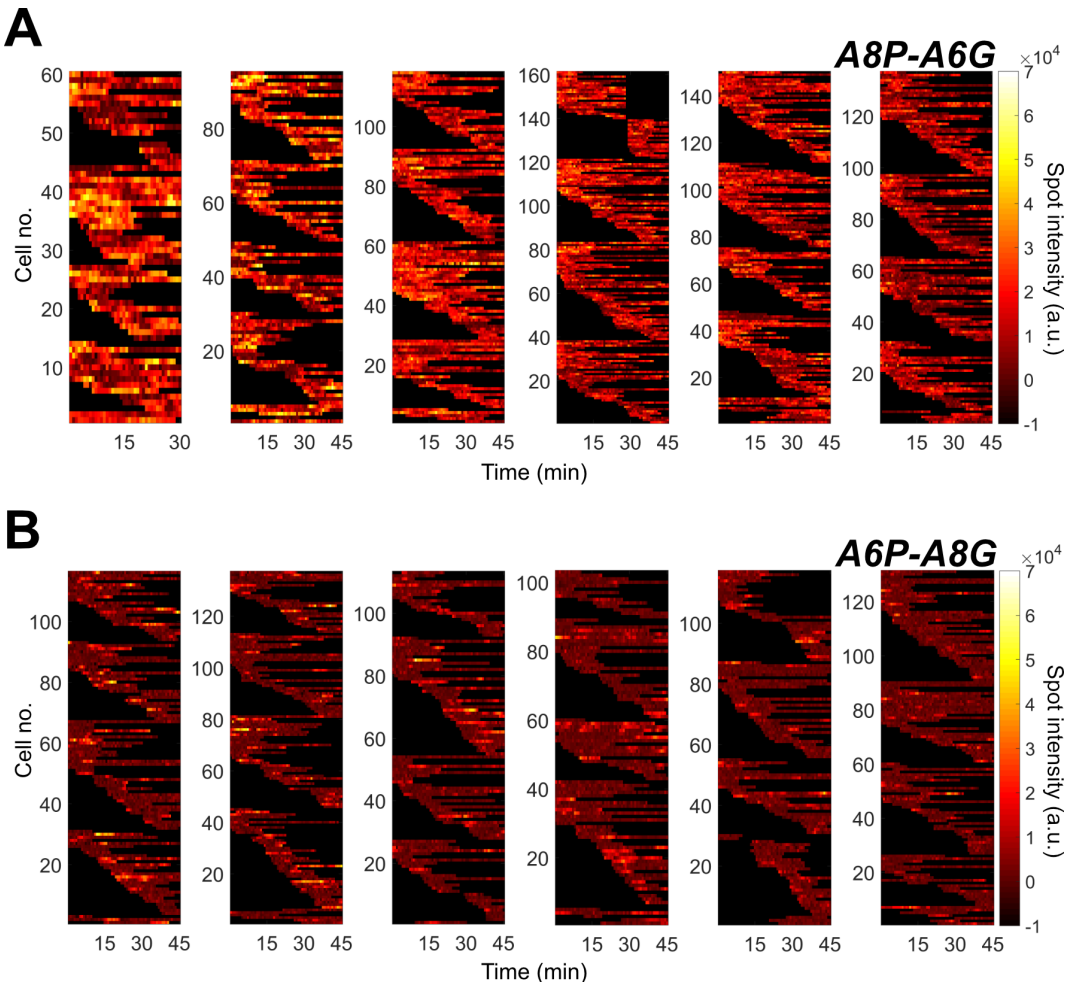

Figure 4.8: Transcription dynamics of actin genes with alternative promoters. Similar to figure 4.4, for the two promoter-switched cell lines - (A) A8P-A6G and (B) A6P-A8G - spot intensity traces for all tracked cells from three experimental days, with two independent clones imaged each day, are shown. As before, colours represent variable gene activity, while black represents no data for the tracked cell in that particular row.

at the endogenous *act8* locus. Similarly, integration of the *act6* promoter at the *act8* gene ('A6P-A8G') reduces the average spot intensity and introduces clear transcriptional bursting behaviour, typical of the endogenous *act6* gene (Fig. 4.8).

Furthermore, the spot intensity distributions resemble those of the endogenous gene to which the modified promoter belongs (Fig. 4.9 A). Statistical comparisons of these distributions proved difficult as even comparisons between experimental repeats for the same gene were statistically

Gene	Median (x 10^4)	Mean (x 10^4)
act5	0.59,0.49,1.1	1.1,1.1,1.6
act6	0.15,0.14,0.22	0.37,0.31,0.57
act8	1.5,2.0,1.9	1.5,1.9,1.8
A8P-A6G	1.7,1.5,1.5,1.2,1.2,0.74	1.7,1.7,1.5,1.3,1.2,0.93
A6P-A8G	0.20,0.19,0.18,0.19,0.20,0.17	0.43,0.41,0.38,0.37,0.33,0.28
Gene	Variance (x 10^8)	Skewness
act5	1.5,1.5,2.6	1.0,1.1,0.73
act6	0.44,0.29,0.85	2.5,2.5,2.0
act8	0.92,0.92,1.4	0.24,-0.23,0.13
A8P-A6G	1.8,1.8,0.96,0.90,0.90,0.69	0.45,0.56,0.41,0.48,0.45,0.77
A6P-A8G	0.59,0.57,0.40,0.33,0.25,0.21	2.4,2.5,2.6,2.2,2.6,2.8

Table 4.1: Summary of statistics describing actin gene spot intensity distributions. Statistics were calculated for all experimental replicates of imaging experiments for all genes. Skewness is calculated as Pearson's moment coefficient of skewness in MATLAB which measures the relative skew of each distribution away from that of a normal distribution.

significant by Kolmogorov-Smirnov and Wilcoxon signed rank tests, due to the large number of data points and the significant experimental variability of the system. Therefore, a summary of the main statistics of all replicate experiments is presented in table 4.1. Comparing these statistics for promoters in their endogenous and switched loci shows that on the whole the distributions are similar, as the range of values for each statistic are similar for both genes. However, the skewness of *act8* appears to be close to zero while A8P-A6G skewness is seemingly higher at around 0.4-0.5. This statistic measures how skewed a particular distribution is away from a normal distribution and therefore A8P-A6G spot intensities are less normally distributed than *act8*. This indicates that while the majority of the data appear to be comparable between genes controlled by the same promoter in two genomic loci, there may be some influence from the local chromosomal environment (or indeed small 'scar' sequences left over from the genetic engineering).

The noise level (C_v^2, $\left(\frac{\sigma}{\mu}\right)^2$) of each promoter-switched gene is sig-

nificantly different (paired Student's t-test; Fig. 4.8 B) as in the endogenous locus of the promoters (Fig. 4.6 A). However, the differences between the noise strength (Fano factor, $\frac{\sigma^2}{\mu}$) of the two genes appears to be reduced (Fig. 4.8 C) while still statistically significant (paired Student's t-test; non-significant using Welch's unequal variance t-test). This could represent differential modulation of bursting parameters in the non-native settings of these promoters, as changes in noise strength (described by the Fano factor) are thought to represent changes in burst size, while changes in noise (defined as the squared coefficient of variation, C_v^2) reflect altered burst frequencies (To & Maheshri, 2010; Carey et al., 2013).

To further characterise the altered dynamic behaviours, I generated cooccurrence matrices for the endogenous and promoter-switched cell lines (Fig. 4.10). This method enables an evaluation of the rate of change of a particular variable over a linear sequence of events. The result of the calculation is a square matrix in which the XY coordinates of each data point represent the spot intensity at time t and t+2 expressed as a percentage of the maximum range of spot intensity values for all live imaging experiments. A lag of two frames was used to avoid capturing noise inherent within single frame transitions as shown previously (Corrigan et al., 2016). Figure 4.10 shows around 8000 frame transitions for both the normal and promoter-switched cell lines from at least 3 independent experiments. A clear signature of transcriptional bursting is visible as the band of data points between 5-15% on both axes. In real terms, this is a representation of those cells in which the spot intensity has rapidly changed from very low values (e.g. S_t = 5-15%) to relatively high values (e.g. S_{t+2} = 20-60% in *act8*) from one frame of the movie to the next, and vice versa. If the gene activity, and therefore spot intensity, stays more constant over time then this will be represented by values closer to the identity line of the matrix.

As expected given the general pattern of transcription in these cell lines the bursting signatures of these genes are largely consistent with those of

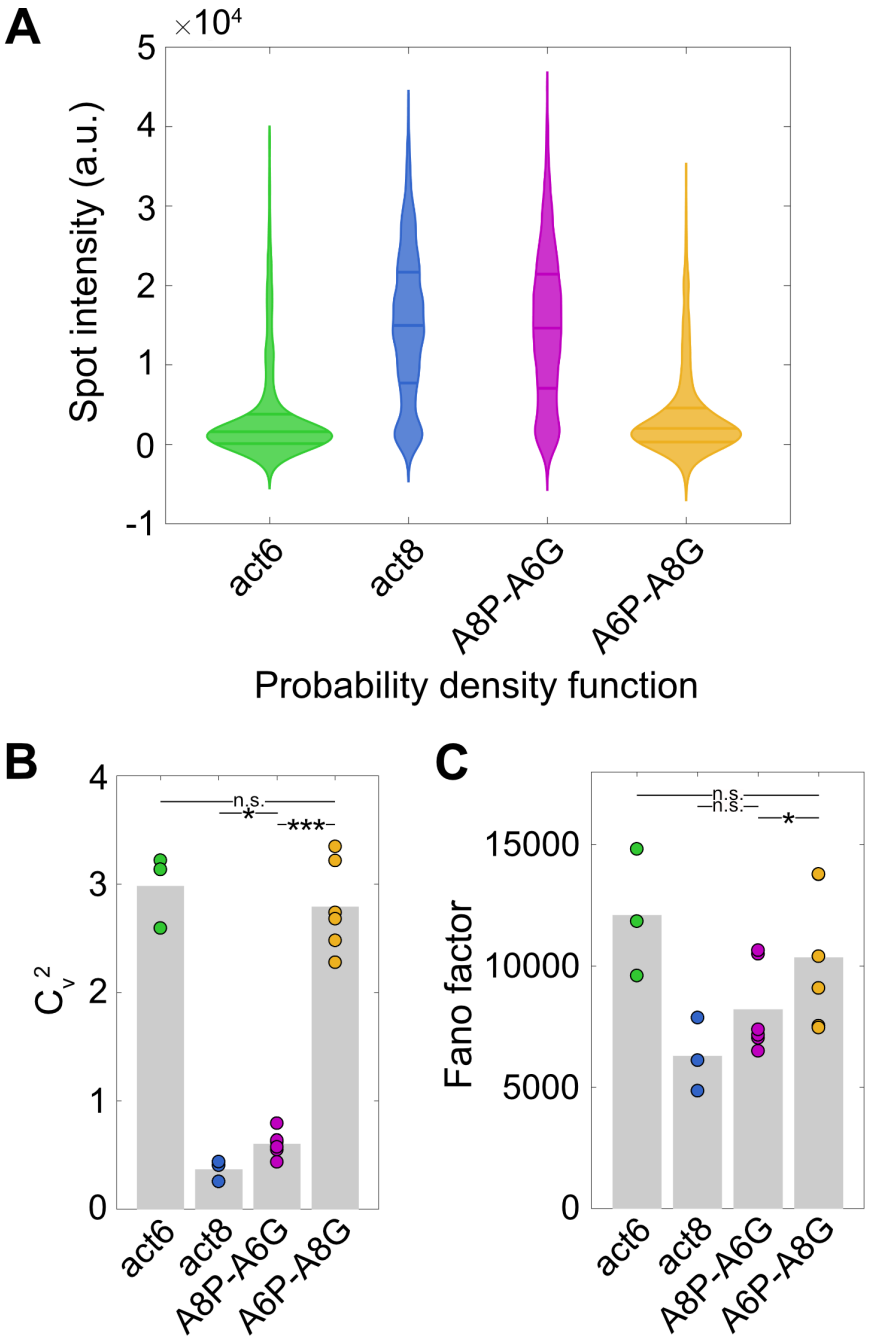

Figure 4.9: Spot intensity distributions of promoter-switched cell lines. (A) Smoothed probability density functions of alternative promoter spot intensities compared to the endogenous loci. (B) Squared coefficient of variation and (C) Fano factor of all spot intensity data, comparing promoter switches with endogenous genes. Grey bars are mean, coloured circles are individual experiments. Significance determined using paired Student's t-test for promoter-switched genes, as these were imaged on the same day, and Welch's (unpaired) unequal variance t-test gave similar results for all other comparisons. N.s. = $p > 0.05$, * = $p < 0.05$, *** = $p < 0.001$.

the gene from which the promoter derives. Most of the spot intensity fluctuations in A8P-A6G involve relatively small changes of intensity, similar to the endogenous *act8* behaviour (Fig. 4.10). Likewise, A6P-A8G intensity fluctuations are more substantial with relatively fewer transitions close to the identity line of the matrix, again similar to *act6* (Fig. 4.10). However, the distinction between these 'bursty' (rapidly changing) and 'non-bursty' (slowly fluctuating) contributions to spot intensity at the promoter-switched genes are less clear than *act6* and *act8* alone. Also, the proportion of slowly fluctuating frame-to-frame transitions in A6P-A8G (data points away from 5-15% bands on both axes) are generally more spread compared to *act6*. These differences could represent minor alterations in the regulation of transcription from these promoters compared to the endogenous genes, but equally could simply be the result of experimental variability.

In summary, it seems that the majority of control of transcription in actin genes comes directly from the promoter sequence, as most elements of gene activity are faithfully recapitulated upon promoter switching. Similar observations of promoter-defined transcriptional bursting have been observed in both fixed (Hocine et al., 2015) and live cells (Yunger et al., 2010) but the above observation is unique in its use of natural promoters at multiple endogenous loci. However, some subtle differences in transcription dynamics of promoters at alternative genomic loci could represent the existence of further minor influences on transcription, such as the local chromatin environment or nuclear organisation.

4.5 Relationships between promoter sequence features and transcriptional dynamics

Having seen that promoter sequence controls transcription in *D. discoideum* actin genes, are there specific features of these sequences which are important for encoding the differential gene regulation? Referring back to the multiple sequence alignment in figure 3.7, the sequence elements highlighted

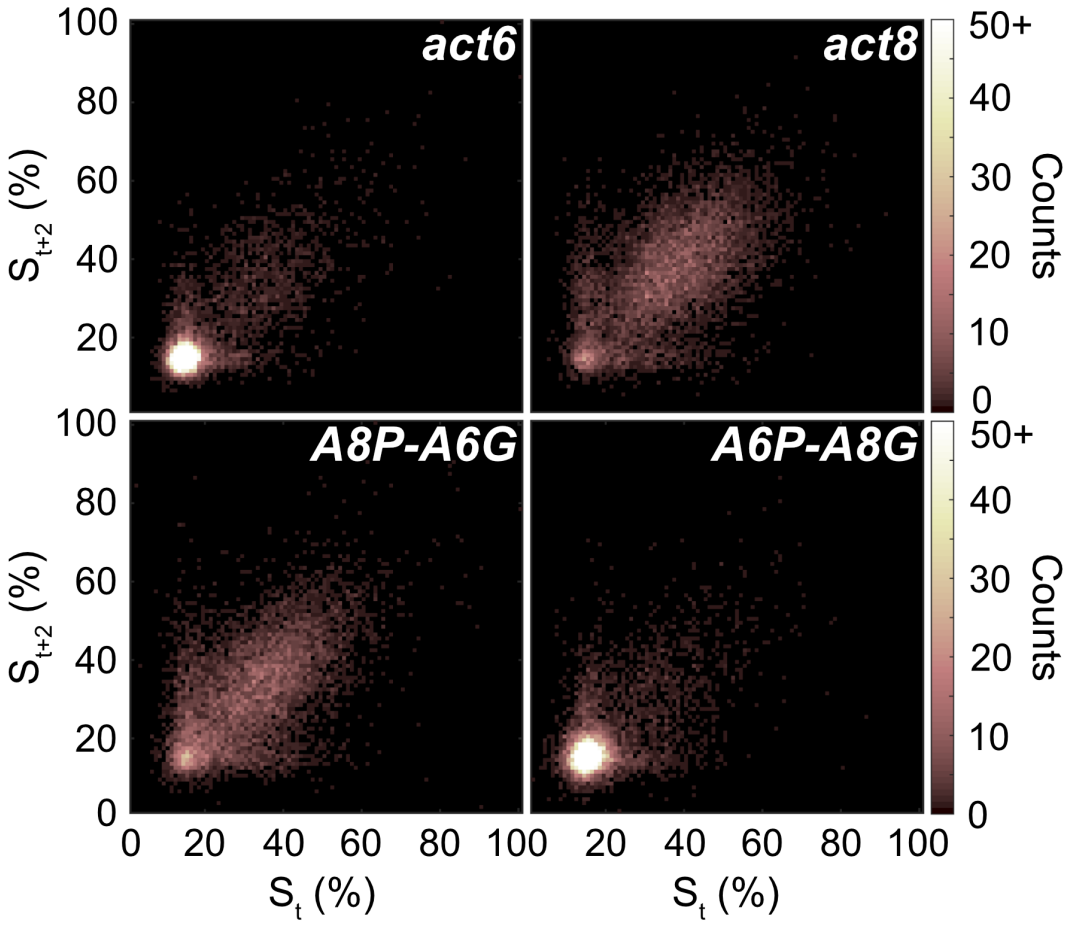

Figure 4.10: Cooccurrence matrices of transcription dynamics from promoter-switched actin genes. Each data point represents a transition between two imaging frames, with the coordinates in X and Y indicating the relative spot intensity for an individual cell before (at time t, S_t) and after the transition (time t+1, S_{t+2}). Relative spot intensity is defined as a percentile of the maximum range of spot intensities across all imaging data. If a relatively weak spot (e.g. 20% of maximum) becomes very bright by the next imaging frame (e.g. 60% of maximum) then this data point would be marked at position (20,60) in the matrix. Around 8000 spot intensity transitions, from three independent experiments, are shown for each cell line to ensure.

there correlate well with the bursting dynamics described in this chapter, according to other studies of transcriptional regulation. Firstly, 'open promoters', or those which enable constitutive, less noisy transcription (Cairns, 2009), typically have reduced nucleosome occupancy (Brown et al., 2013; Dey et al., 2015) which in turn is associated with the presence of poly(dA:dT) tracts (Segal & Widom, 2009). Separate observations show that poly(dA:dT) stretches are also correlated with high levels of low noise gene expression (Struhl, 1985; Sharon et al., 2012, 2014). Therefore, the fact that *act8* has the highest proportion of its core promoter made up of long ($>$10) runs of poly(dA:dT) of all the actin genes (Fig. 3.7) marries well with the observation of high levels of transcription with low variability (Fig. 4.5). *act5* and *act6* by comparison have a much smaller proportion of their promoters made of nucleosome disfavouring sequences (Fig. 3.7) and also exhibit noisier transcription.

The presence of a TATA box has been shown on numerous occasions to facilitate noisy gene expression (Raser & O'Shea, 2004; Blake et al., 2006; Tirosh & Barkai, 2008; Hornung et al., 2012). The two genes in this analysis exhibiting noisy gene expression, *act5* and *act6*, both have a canonical TATA box about 30 bp upstream of the predicted TSS, while *act8*, with much less noisy transcription, does not (Figs. 3.7, 4.5). Indeed, a detailed analysis of the mechanisms controlling *act5* transcription showed that the TATA box was important for accessing the very highest initiation rates (Corrigan et al., 2016). Here, the maximal *act5* spot intensities are higher even than the strongest gene activity of the TATA-less *act8* gene (Fig. 4.5), despite the mean spot intensity of *act5* being less than *act8*.

Other promoter sequence-specific determinants of transcriptional variability include complexity of the promoter in terms of the number of transcription factor (TF) binding sites. Use of the tet-Off system showed that reporter genes with 7 tetO binding sites, as opposed to one, showed more noise in gene expression (Raj et al., 2006; To & Maheshri, 2010) while similar results

were observed using a library of hundreds of designed promoter sequences with variable numbers of Gcn4 binding sites in yeast (Sharon et al., 2014). Although the multiple sequence alignment in figure 3.7 does not explicitly reveal transcription factor binding sites, the high level of sequence conservation, strong dyad symmetry (Khorasanizadeh & Rastinejad, 2001) and previous inference of importance for gene expression (Hori & Firtel, 1994) would suggest that at least some of the motifs identified are bound by TFs. Assuming this is the case, then the fact that *act5* and in particular *act6* harbour many more putative TF binding sites in their promoters compared to *act8* (Fig. 3.7) also aligns with published literature in terms of the effects of promoter sequence on noisy transcription. The single binding site that is present in the *act8* promoter is a CCAAT box, which has been shown to increase the fraction of time a gene is on by decreasing the wait times between transcriptional bursts (Suter et al., 2011). Again, this appears to tie in with the transcriptional behaviours outlined previously (Fig. 4.4).

Overall, the dynamic behaviours of the different actin genes surveyed here are likely controlled by specific promoter components which have been previously described as determinants of various features of gene expression. The seemingly predictive nature of these features could enable accurate prediction of the dynamic behaviour of other genes within the family.

4.6 Effects of extrinsic variability on actin gene activity

Transcription of a gene is often the end point of upstream signalling processes, with cells constantly sampling their environments and coordinating gene expression changes accordingly (Brivanlou & Darnell, 2002; Weake & Workman, 2010; White & Sharrocks, 2010). I have established that the differences in transcription dynamics of different actin genes are likely due to promoter sequence variability, but what are the regulatory inputs upstream of these diverse transcriptional events? To explore the ability of diverse

sources of variability to influence actin gene expression I utilised the many additional parameters, besides spot intensity, measured during the segmentation and tracking analysis of the live imaging movies (Corrigan et al., 2016).

4.6.1 Cell size and actin gene activity

Firstly, I measured the effect of cell size on the activity of each actin gene. The custom-built analysis software used to track cells cannot accurately measure volume (30% error rate) but does measure the 'area' of a cell in two dimensions. While this is not the most accurate descriptor of cell size, by averaging this 'cellular footprint' over the course of an imaging session the relative size of a cell should be captured, despite the extensive exploration of three-dimensional space which *Dictyostelium* cells undergo during motility. Figure 4.11 shows the relationship between this averaged cell size measurement and the average spot intensity for all tracked cells. All three actin genes show weak, but significant, correlations between gene activity and cell size (Spearman's rank correlation, *act5*: r = 0.38, p = 0; *act6*: r = 0.29, p = 1.4×10^{-8}; *act8*: r = 0.25, p = 2.4×10^{-5}). Most genes exhibit size-dependent scaling of mRNA count (Zhurinsky et al., 2010; Kempe et al., 2015; Padovan-Merhar et al., 2015), which is due to a global increase transcriptional activity in larger cells (Schmidt & Schibler, 1995; Zhurinsky et al., 2010; Padovan-Merhar et al., 2015). The relatively weak correlations here could be derived from the use of area rather than volume measurements, or the fact that spot intensity, and therefore gene activity, is only measured for about 20 min per cell on average which may not exactly mirror the cytoplasmic mRNA content of the cell (quantitative measures of mRNA content are used later in this chapter). Equally, not all cell-to-cell variability can be explained by size alone (Padovan-Merhar et al., 2015) and other factors are therefore likely to be involved, of which stochasticity in gene expression is likely one.

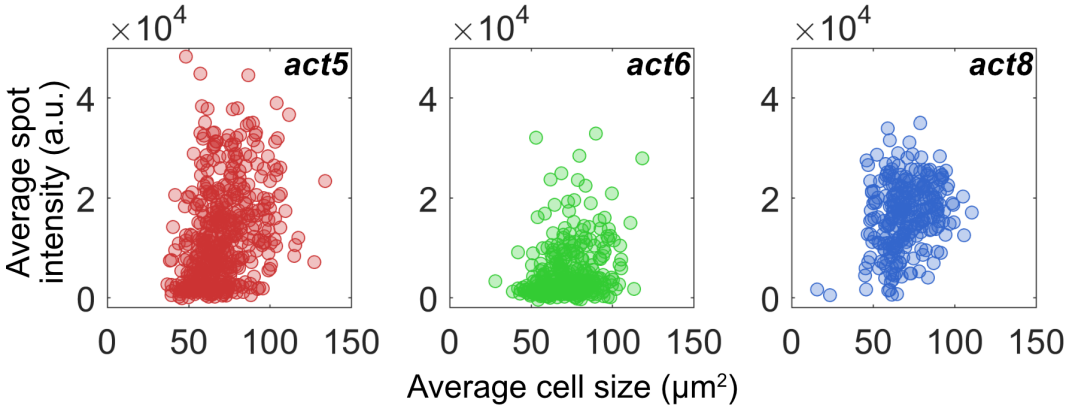

Figure 4.11: Cell size correlations with actin gene spot intensity. For each tracked cell, spot intensity values and cellular 'footprint' were averaged over the movie and correlated. Spearman's rank correlation, *act5*: r = 0.38, p = 0, n = 478; *act6*: r = 0.29, p = 1.4×10^{-08}, n = 384; *act8*: r = 0.25, p = 2.4×10^{-05}, n = 275.

4.6.2 Population density and actin gene activity

Next, I measured the effect of population density on spot intensity. To do this, I calculated the average number of cells in the imaging frame over the duration of the movie for individual fields of view (FOV). I then calculated the average spot intensity for each FOV and correlated the two for each actin gene (Fig. 4.12 A). Despite the relatively few data points and the differences in density on different experimental days (signified by different shades of each colour), there appears to be a relationship between population density and average spot intensity for *act5* (Spearman's rank correlation, r = -0.62, p = 0.012) and *act6* (r = - 0.80, p = 0.0003) but not for *act8* (r = -0.20, p = 0.46). This could suggest that highly variable actin genes are more responsive to changes in the density of a population of cells, with increased rates of transcription when the population is sparse and reduced transcription when the population is dense. Studies in other systems have shown similar results (McKinnon & Burgoyne, 1984; Schmitt-Ney & Habener, 2004) although this has been contradicted elsewhere (Greer et al., 2010). Recent work has demonstrated the mechanism by which *Dictyostelium* sense culture density (Suess & Gomer, 2016) and shown that this extracellular signal reduces the levels of actin cytoskeleton proteins via Ras and Akt (Suess et al., 2017),

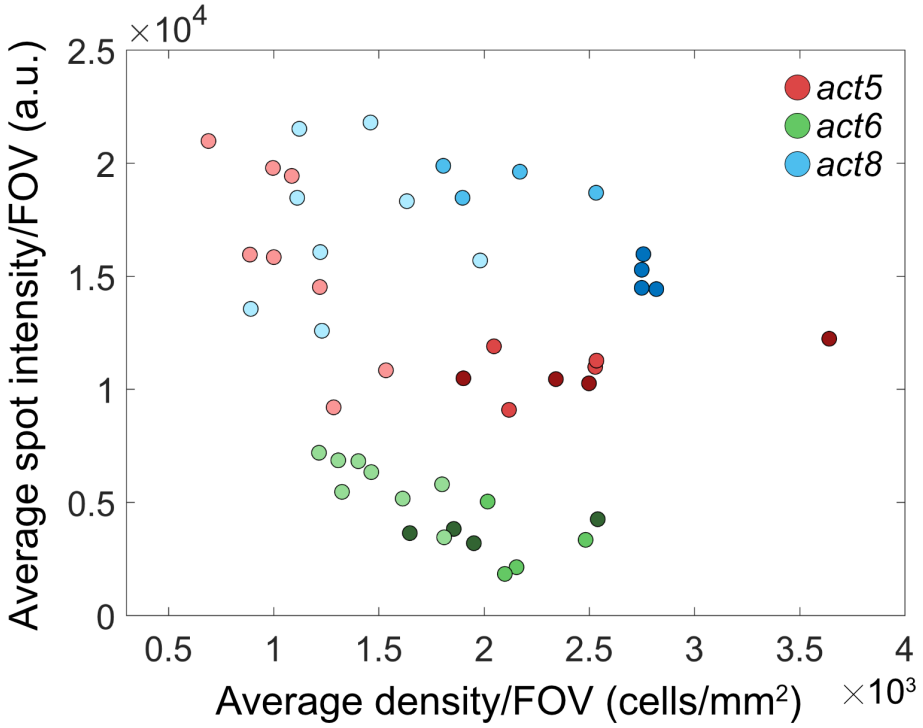

Figure 4.12: Population density correlations with actin gene spot intensity. Movies were separated into fields of view (FOV) and density calculated as the number of nuclei in a single frame. Average spot intensity was correlated with the average density per field of view over the course of an imaging session. Shades of each colour represent different experimental days. Spearman's rank correlation, *act5*: r = -0.62, p = 0.012; *act6*: r = - 0.80, p = 0.0003; *act8*: r = -0.20, p = 0.46.

lending strong support to the results shown above. This type of differential regulation in response to environmental cues could be an important justification for expansion of the gene family.

4.6.3 Cell motility and actin gene activity

Cell motility is intricately linked to the actin cytoskeleton (Pollard & Borisy, 2003; Blanchoin et al., 2014) while *Dictyostelium* are a highly motile cell type. I assessed whether actin gene activity is associated with cell motility by calculating the average speed of individual cells while imaging transcriptional dynamics. Figure 4.13 shows the correlations of these two parameters. As with cell size, all three genes show weak but significant correlations between individual cell speed and spot intensity (Spearman's rank correla-

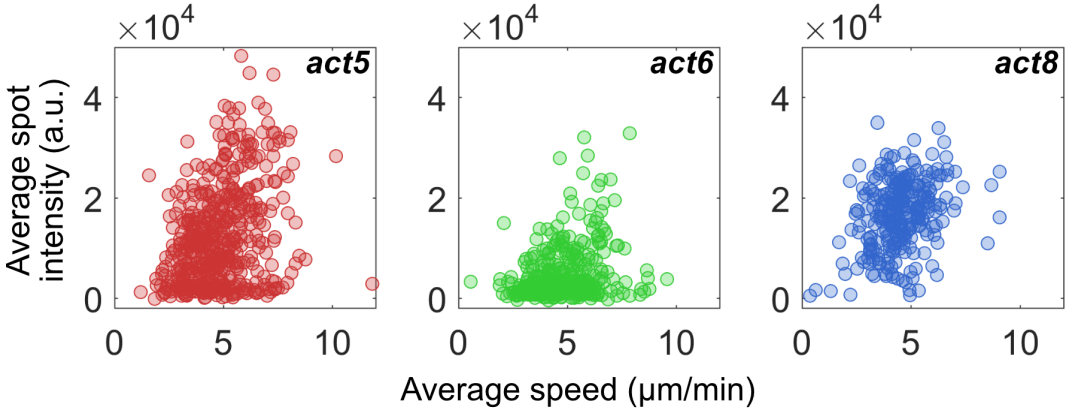

Figure 4.13: Cell motility correlations with actin gene spot intensity. For each tracked cell, spot intensity and cell speed were averaged over the movie and correlated. Spearman's rank correlation, *act5*: r = 0.32, p = 1.5×10^{-12}, n = 478; *act6*: r = 0.26, p = 3.0×10^{-7}, n = 384; *act8*: r = 0.29, p = 1.4×10^{-6}, n = 275.

tion, *act5*: r = 0.32, p = 1.5×10^{-12}; *act6*: r = 0.26, p = 3.0×10^{-7}; *act8*: r = 0.29, p = 1.4×10^{-6}). This implies that some of the variability in spot intensity is accounted for by differences in the speed of cells, with faster cells more likely to have strongly active actin genes.

Intuitively, one would assume that the speed of a motile cell is inversely proportional to the amount of free space it has in which to move, which can be otherwise interpreted as the local cell density in an experimental situation. Therefore, as we see an effect on actin gene activity from both culture density and cell motility, I checked how these two variables are correlated in these cells. However, while both *act5* and *act6* show moderate negative correlation between density and speed (Fig. 4.14; Spearman's rank correlation, *act5*: r = -0.55, p = 0.030; *act6*: r = -0.59, p = 0.020) as might be expected, *act8* shows no significant correlation between the two variables (r = 0.42, p = 0.10). A negative correlation between speed and density has been demonstrated previously in *D. discoideum* (Golé et al., 2011), and therefore the fact that cells with MS2-tagged *act8* don't follow this trend suggests that density-dependent motility is affected by this genetic modification in these cells. The MS2-*bsr* targeting construct disrupts the coding sequence at the 5' end (see Fig. 4.1). Therefore, the effective loss of the highly expressed

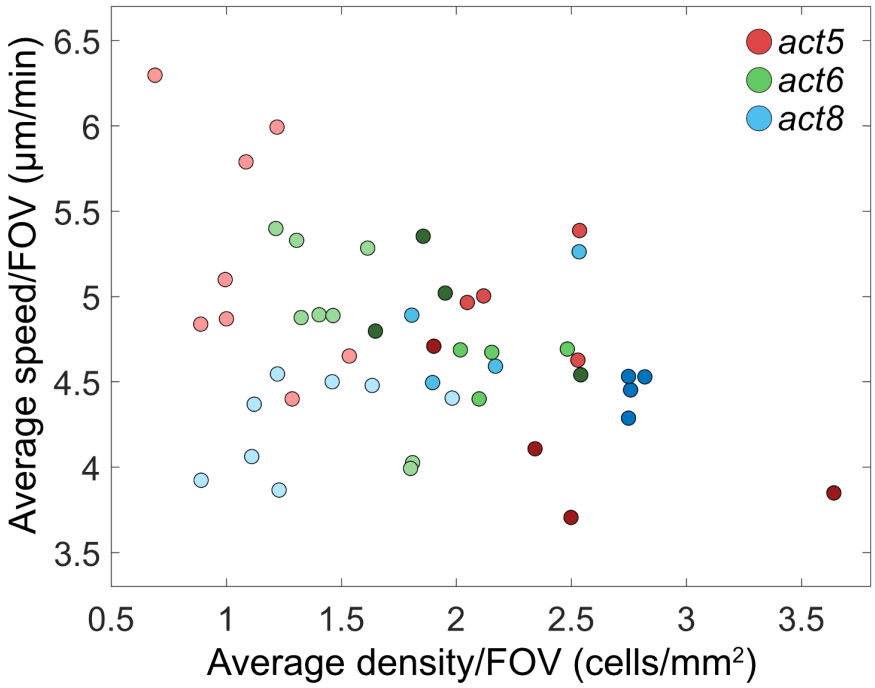

Figure 4.14: Population density correlations with cell motility. Average density and average speed (as calculated previously) were correlated for each field of view. Shades of each colour represent different experimental days. Spearman's rank correlation, *act5*: r = -0.55, p = 0.030; *act6*: r = -0.59, p = 0.020; *act8*: r = 0.42, p = 0.10.

act8 gene in these cells could account for the apparent differences between the responsiveness of actin genes to population density shown in figure 4.12 (further characterisation of actin gene knockouts is described in Chapter 5). It is interesting to note that there is still a significant correlation between speed and transcriptional activity in *act8*-MS2 cells despite the apparent uncoupling of the motile response to population density. This could suggest a requirement for more actin in fast-moving cells, independent of that determined by population density.

In summary, it appears that *D. discoideum* actin gene expression is influenced by numerous regulatory inputs, as described in other systems (McKinnon & Burgoyne, 1984; Schmitt-Ney & Habener, 2004; Padovan-Merhar et al., 2015). Transcriptional activity is only weakly correlated with cell size and speed, while population density is more strongly anti-correlated with *act5* and *act6* but not with *act8*. This could represent differential re-

sponsiveness of actin genes to environmental signals, which would be an important demonstration of differing roles for actin genes within the organism. However, this observation may be confounded by the fact that the MS2-tagged cell lines do not respond in the same way to population density, in terms of cell motility. Further experimentation is required to thoroughly test these ideas.

4.7 Cytoplasmic mRNA counts of *D. discoideum* actin genes

Until now all the data presented in this chapter have been concerned with measurements of instantaneous transcriptional activity. These experiments allow exploration of dynamic behaviours of genes and can potentially identify different immediate responses to stimuli. However, the effectors of most cellular responses are proteins, the production of which requires several steps beyond transcription. To test whether differences in transcription dynamics are visible at later stages of actin protein production I quantified the number of cytoplasmic mRNAs in cells using single molecule RNA fluorescence *in situ* hybridisation (RNA FISH) (Lawrence et al., 1989; Femino et al., 1998).

By using a set of fluorescent probes which hybridise specifically to the MS2 sequence of the transgene (Fig. 4.15 A), single molecules of mRNA derived from particular actin genes can be identified within the cell (Fig. 4.15 B-E). This technique reveals a more static, steady-state view of gene expression, but highlights similar trends to those seen from live imaging. Qualitatively, *act8* (Fig. 4.15 E) has the most mRNAs per cell on average, followed by *act5* (Fig. 4.15 C) and *act6* (Fig. 4.15 D), similar to the trends in figure 4.5. Interestingly, while MS2-tagged *act1* cells showed no evidence of active transcription in living cells (Fig. 4.2), use of RNA FISH highlights several copies of *act1* mRNA per cell on average (Fig. 4.15 B) demonstrating the increased sensitivity of this method.

To assess the distributions of mature actin transcripts I counted cyto-

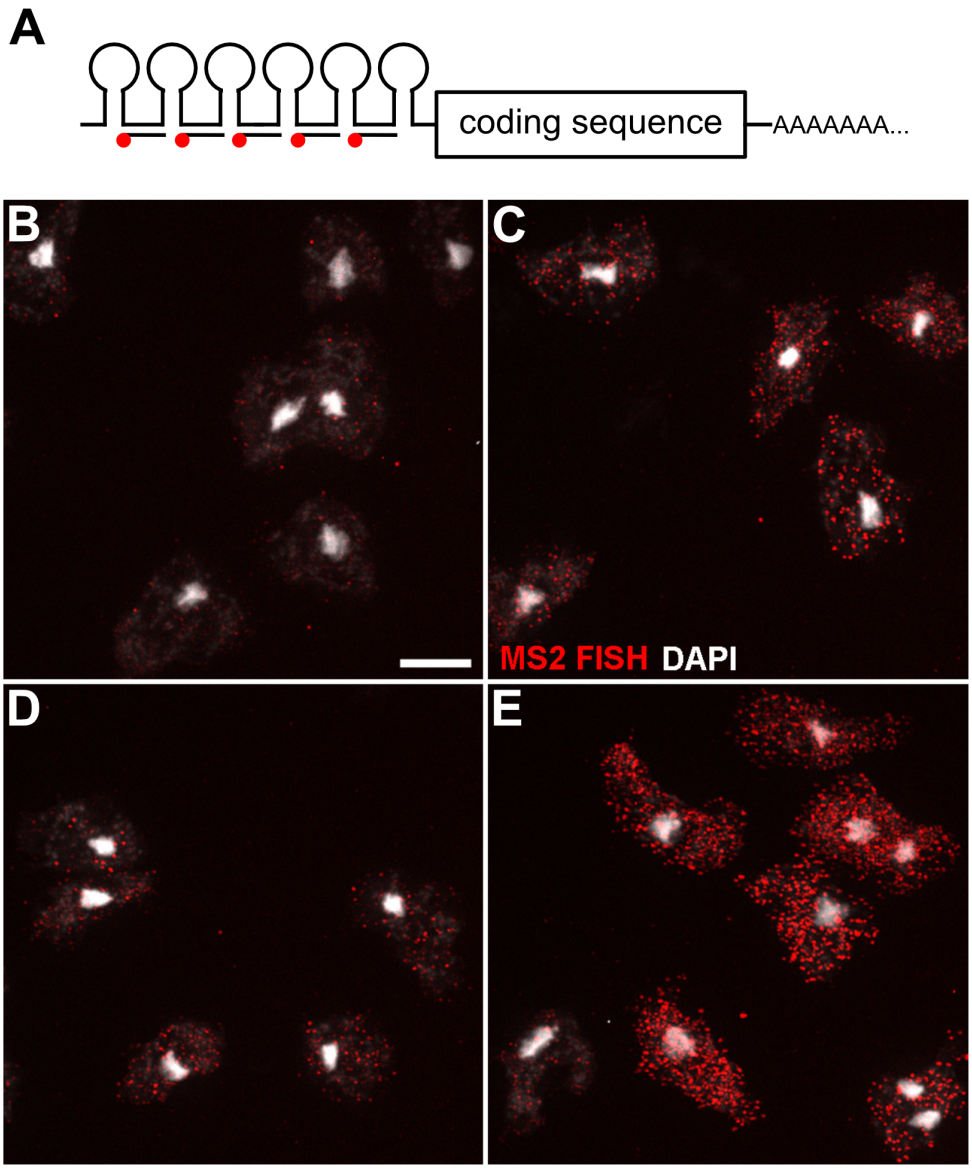

Figure 4.15: mRNA counting by single molecule RNA FISH. (A) RNA FISH probes specific to the MS2 repeats were used to specifically localise transcripts from individual actin genes. mRNAs from (B) *act1*, (C) *act5*, (D) *act6* and (E) *act8* genes were identified in single cells. Bright puncta represent individual transcripts. Scale = 7 μm.

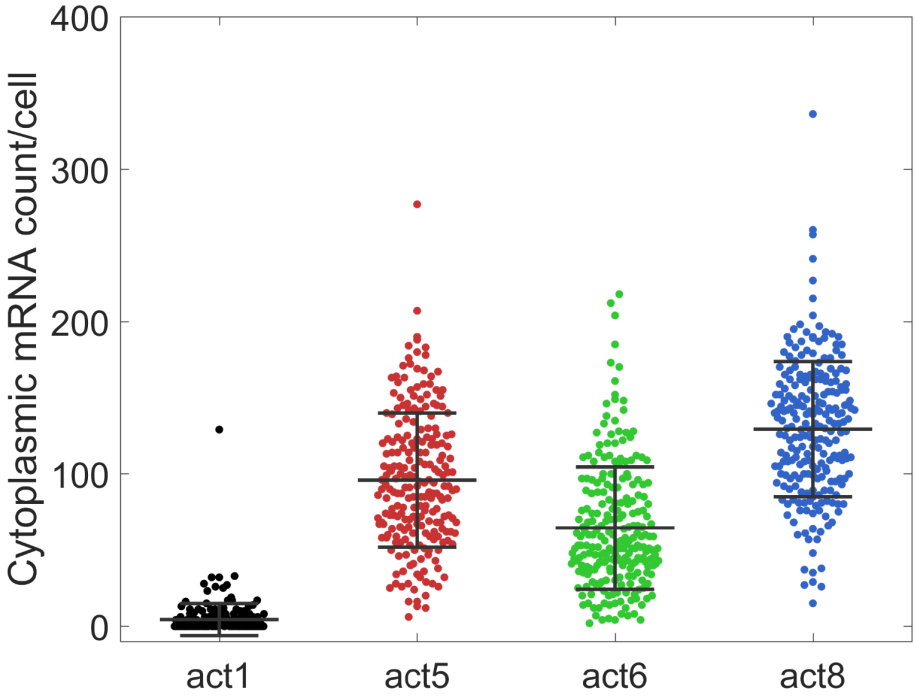

Figure 4.16: Cytoplasmic mRNA distributions of *D. discoideum* actin genes. Cytoplasmic mRNA counts were determined for individual cells using FISH-quant. Data from three experiments are pooled. *act1* n = 217, *act5* n = 225, *act6* n = 239, *act8* n = 238. Bars show mean and standard deviation.

plasmic mRNAs in single cells for all four actin genes using FISH-quant (Mueller et al., 2013) (Fig. 4.16). In comparison to the distributions of spot intensities described earlier (Fig. 4.5) the spread of *act5* and *act6* counts are more normally distributed and much less skewed towards zero (Fig. 4.16). The mean values of *act5* and *act6* are also much closer to *act8* compared to the live imaging dataset. As mentioned, while most cells have none or only a few *act1* transcripts per cell, a small proportion have several tens of transcripts, while one has over 100 (Fig. 4.16).

To probe this further I measured the variability of the highly expressed actin genes tested here (Fig. 4.17). The pattern of noise in the system reflects that found at the transcriptional level (Fig. 4.6) with *act8* the least noisy, followed by *act5* and *act6* (Fig. 4.17 A). However, the magnitude of the differences between *act8* noise and that of the other actin genes is reduced compared to nascent transcription dynamics (Figs. 4.6 A, 4.17 A).

Indeed, *act5* and *act8* noise (C_v^2) are not significantly different (paired Student's t-test, $p < 0.05$ using Welch's unpaired t-test; Fig. 4.17 A) while noise strength (Fano factor) values are indistinguishable from each other (paired Student's t-test), where *act8* was significantly different at the transcriptional level (Figs. 4.6 B, 4.17 B). It is also important to note here that the ratio of the variance to the mean number of mRNA counts per cell is typically used as a gauge of whether gene expression follows a Poisson distribution or not (Raj et al., 2006). If this value is close to 1 then a Poisson distribution can be inferred; clearly none of these genes are produced via a one-state Poisson process (Fig. 4.17 B) and therefore all three genes, including *act8* are likely transcribed with bursting behaviours. On the whole, the general trends of these data are in agreement with live imaging of transcription, but the shape and variability of the distributions of mature mRNAs are different, perhaps due to the extended lifetimes of these molecules relative to the dynamic nature of transcriptional states (Corrigan et al., 2016). This calls into question whether these bursting dynamics are functionally relevant to protein production.

4.8 Population distributions of mRNA in promoter-switched cell lines

Having seen that transcription dynamics of promoter-switched cell lines are near-identical to that of the endogenous gene I measured how this symmetry translates to the distribution of cytoplasmic mRNAs in a population of single cells. Figure 4.18 shows that, as with the transcription dynamics, the distribution and variability of cytoplasmic mRNAs in promoter-switched cell lines largely mirror those of the endogenous gene locus; the *act8* promoter generates less noisy steady-state mRNA distributions compared to *act6* promoter (paired Student's t-test; Fig. 4.18), although in this experiment the noise strength was also significantly different which wasn't the case with promoters at endogenous loci (Figs. 4.17 B, 4.18 C).

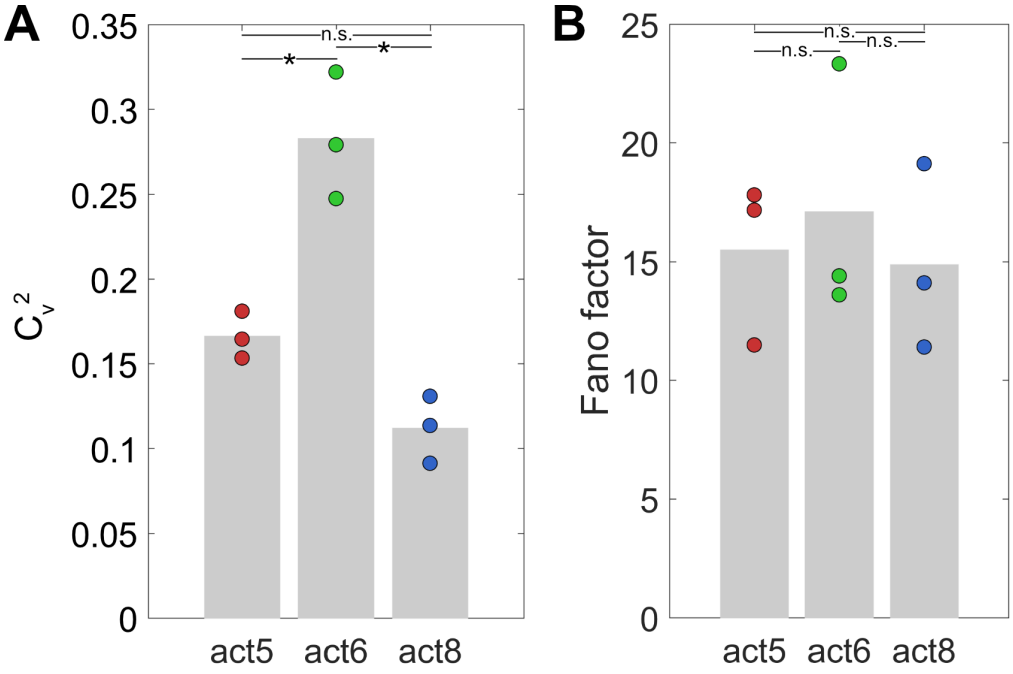

Figure 4.17: Variability of actin mRNA distributions. (A) Squared coefficient of variation $\left(\frac{\sigma}{\mu}\right)^2$ and (B) Fano factor ($\frac{\sigma^2}{\mu}$) were calculated for all RNA FISH data from three experiments. Mean values are grey bars, while individual data points are coloured circles. Significance determined using paired Student's t-test (Welch's (unpaired) unequal variance t-test gave similar results unless stated). N.s. = p >0.05, * = p <0.05.

One observation made during the analysis of FISH datasets was the presence of what appeared to be RNA aggregates in the 'A8P-A6G' cell line, which weren't present in endogenous *act6* or *act8* cells. Examples of these granule-like structures, found in two independent clones, can be seen in figure 4.19 (black arrowheads). Interestingly, this type of structure can also be found in *act5* cells, and therefore it could be that specific interactions between 5' and 3' UTR, made possible by abnormal promoter context in the A8P-A6G cell line, facilitate their formation.

4.9 Cell size relationship to cytoplasmic mRNA counts

In section 4.6.1 I showed that transcriptional activity of actin genes is only weakly correlated with cell size, potentially as a result of monitoring such activity for a relatively short period of time. How do these analyses compare

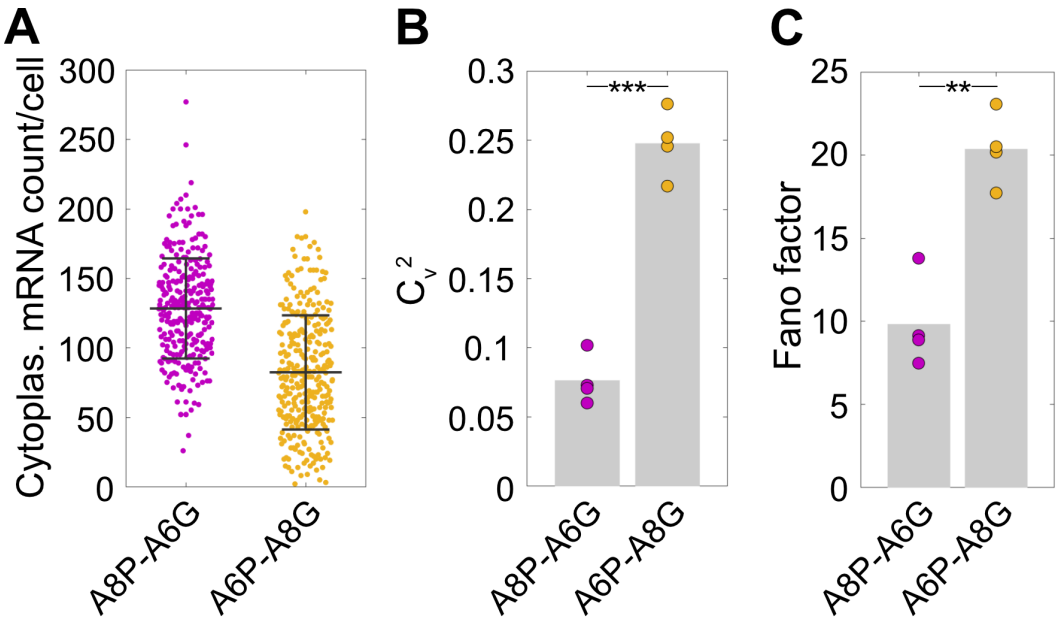

Figure 4.18: Distributions and variability of cytoplasmic mRNAs in promoter-switched cell populations. (A) mRNA counts for promoter-switched cell lines were determined with FISH-quant. A8P-A6G n = 292, A6P-A8G n = 321. Bars represent mean and standard deviation. (B) Squared coefficient of variation $\left(\frac{\sigma}{\mu}\right)^2$ and (C) Fano factor ($\frac{\sigma^2}{\mu}$) were calculated for all data from four experiments. Mean values are grey bars, while individual data points are coloured circles. Significance determined using paired Student's t-test (Welch's (unpaired) unequal variance t-test gave similar results unless stated). ** = p $<$0.01, *** = p $<$0.001.

with 'steady-state' distributions of mRNA? Figure 4.20 shows correlations between cell size and cytoplasmic mRNA counts for different actin genes. As before, cell size is defined as the 'footprint' of the cell in the maximum projection image used to quantify mRNA counts in FISH-quant, and is therefore unlikely to be as accurate as an exact measure of volume. Actin mRNA counts are more strongly correlated with cell size than average transcriptional activities as described previously (Figs. 4.11, 4.20 A) (Spearman's rank correlation, *act1*: r = 0.25, p = 2.2×10^{-4}; *act5*: r = 0.57, p = 5.2×10^{-21}; *act6*: r = 0.48, p = 7.0×10^{-15}; *act8*: r = 0.64, p = 2.8×10^{-29}) which is in line with previous data showing a significant size-dependent component to actin mRNA variability (Padovan-Merhar et al., 2015). However, cell size does not explain all of the variability within a population, given the sizeable range of

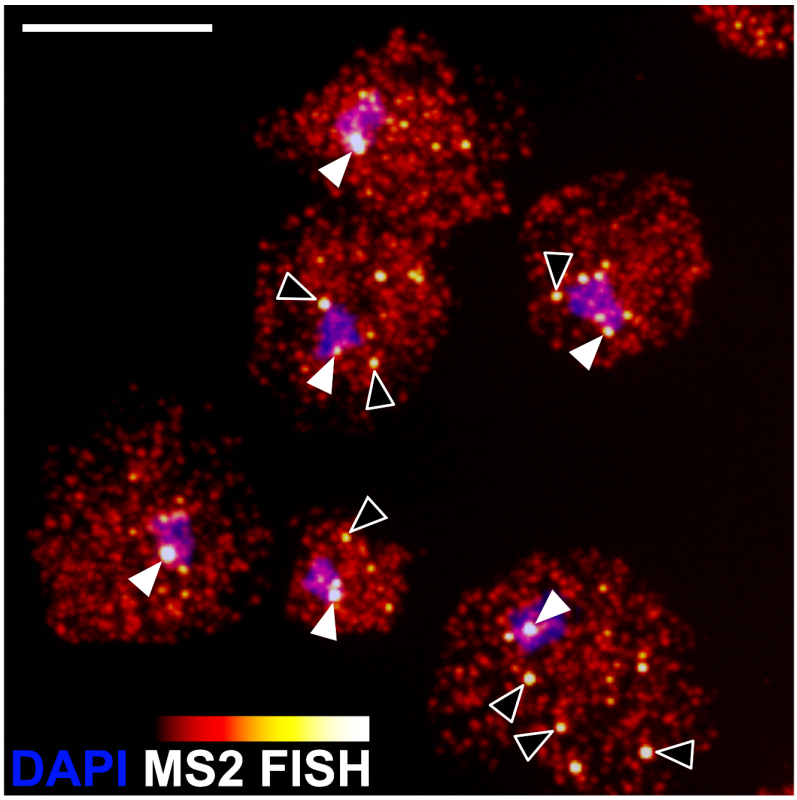

Figure 4.19: Unusual RNA granule-like structures in promoter-switched cell lines. Several high intensity can be seen in A8P-A6G cells in RNA FISH experiments. Transcription sites (white arrowheads) are analogous to those seen in live imaging experiments. Bright cytoplasmic structures (black arrowheads) are not seen in either *act6* or *act8* cells but are sometimes found in *act5* experiments. Scale bar = 10μm

mRNA counts for a particular cell area (Fig. 4.20).

To determine how the scaling of mRNA count changes with size for the different actin genes I calculated the gradient of the linear regression for *act5*, *act6* and *act8* (Fig. 4.20 B). The rate at which mRNA count scales with cell size is significantly higher for *act8* compared to *act5* (paired Student's t-test, p = 1.3×10^{-5}). The gradients of *act6* and *act8* are not significantly different (despite a lower mean value for *act6* compared to *act5*) when using the paired Student's t-test (p = 0.0955) due to the fact that highest *act6* and lowest *act8* data point are paired. Use of unpaired Welch's (unequal variance) t-test gives significance at 5% (p = 0.0373). Overall, this could suggest other more subtle features of the relationship between cell size and

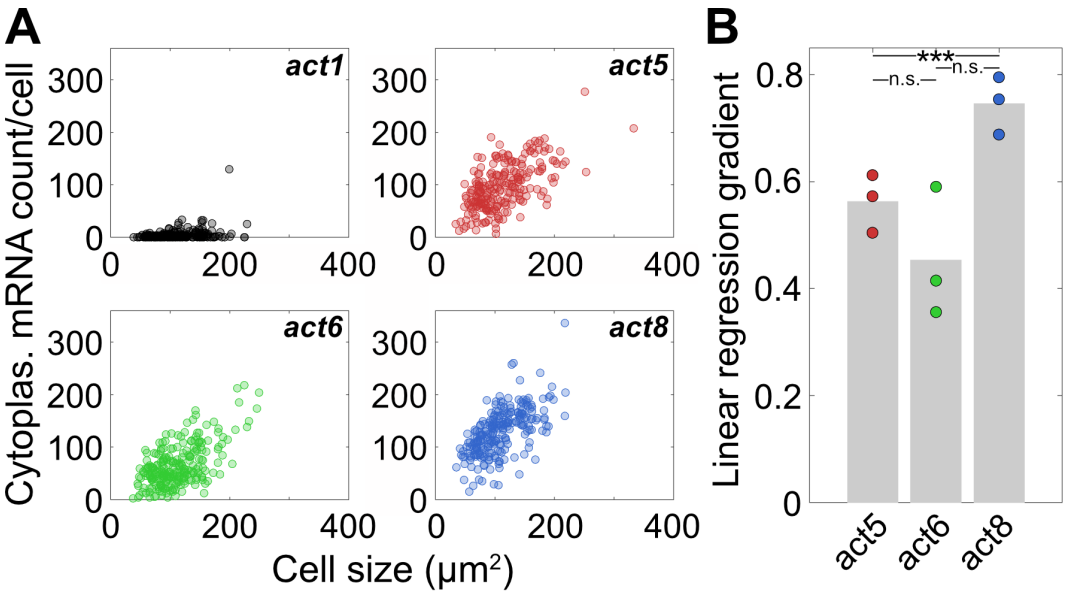

Figure 4.20: Relationship between cell size and cytoplasmic mRNA count. (A) Scatter plots correlating cell size with cytoplasmic mRNA counts for different actin genes. Spearman's rank correlation, *act1*: r = 0.25, p = 2.2×10^{-4}; *act5*: r = 0.57, p = 5.2×10^{-21}; *act6*: r = 0.48, p = 7.0×10^{-15}; *act8*: r = 0.64, p = 2.8×10^{-29}. (B) Gradients of the linear regression from three separate experiments are shown. Grey bars are means, coloured circles are individual experiments. Significance determined using paired Student's t-test (Welch's (unpaired) unequal variance t-test gave similar results). N.s. = p $>$0.05, *** = p $<$0.001. [N.B.: The increased size of cells in this experiment compared to figure 4.11 are likely to be due to the use of fixatives in the FISH protocol which, anecdotally, seem to flatten cells with increasing incubation time.]

actin gene expression. *act8* mRNA count being more closely tied to cell size could be consistent with a role for this gene in basal production of the protein, as opposed to dynamic responsiveness of other genes. Further experimentation would be required to test this.

4.10 Noise reduction mechanisms in actin gene expression

The difference in noise between *act8*, which is consistently transcribed at high levels, and *act5* (and *act6*) which exhibits significant 'off' periods between transcriptional bursts, is significant when comparing transcription dynamics (Fig. 4.6) but reduced in cytoplasmic mRNA distributions (Fig. 4.17). To explore potential mechanisms behind this reduction I first compared the

distributions of cytoplasmic and nuclear mRNAs in cells, as regulated nuclear export of mRNA has been posited as a mechanism for noise reduction in gene expression (Battich et al., 2015; Bahar Halpern et al., 2015a).

Figure 4.21 A shows scatter plots of nuclear and cytoplasmic mRNAs in single cells for all four actin genes. If noise reduction in *act5*, for example, is facilitated by delayed nuclear export then one might expect to see a difference in the ratio of nuclear to cytoplasmic mRNAs compared to *act8*. In figure 4.21 A this would manifest as changes in the slope of regression between the two variables. Figure 4.21 B shows the gradients of linear regression fitting for the different actin genes. *act5*, *act6* and *act8* exhibit very similar relationships between nuclear and cytoplasmic mRNA distributions (except when comparing *act5* and *act6*: p = 0.021 using paired Student's t-test, p = 0.62 using unpaired Student's t-test) suggesting that nuclear export of these mRNAs is unlikely to be differentially regulated. However, the gradient of the linear fit between *act1* nuclear and cytoplasmic mRNAs is significantly higher than the other actins (Fig. 4.21) and therefore this could suggest a role for the process for this gene.

Since a reduction in noise from nascent mRNA production to cytoplasmic mRNA distributions between *act5* and *act8* does not seem to be regulated by nuclear export I next measured the stability of the mRNAs. Longer lifetimes of *act5* mRNA could reduce variability within a population as the noise from transcriptional bursts would be integrated over a much longer time period. To measure mRNA stability, I harvested RNA from cells during a time course of actinomycin D treatment. Actinomycin D blocks transcription of mRNA in *Dictyostelium* (Firtel et al., 1973) and so by measuring the relative levels of actin mRNAs over such a time course one can generate a decay curve for the transcript. Figure 4.22 shows decay curves from a single experiment for different actin genes. While this experiment needs to be repeated, the data presented here appear to show some interesting trends. Firstly, *act5* and *act6* show faster turnover, and therefore reduced stability,

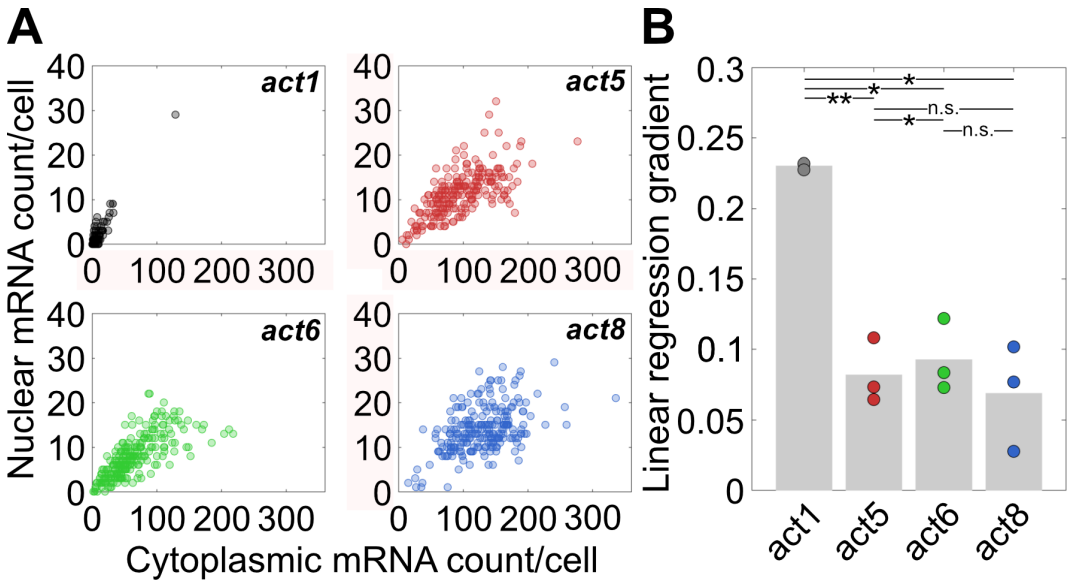

Figure 4.21: Relationship between cytoplasmic and nuclear actin mRNAs. (A) Scatter plots correlating nuclear and cytoplasmic mRNAs. (B) Slope of the linear regression of the relationship described in (A). Grey bars are mean, coloured circles are individual experiments. Significance determined by paired Student's t-test (Welch's unpaired, unequal variance t-test gave similar results unless discussed in text). N.s. = p >0.05, * = p <0.05, ** = p <0.01.

compared to *act8* (Fig. 4.22). If differences in mRNA stability explained the reduction in noise level between genes in cytoplasmic mRNA populations then one would expect the opposite of this result, as more stable *act5* and *act6* transcripts would reduce the heterogeneity between cells introduced during the transcription process. Perhaps even more striking is that the promoter appears to control the relative stability of these transcripts. A6P-A8G is more similar to *act6* while A8P-A6G is more similar to *act8* in terms of the rate of decay (Fig. 4.22). This is surprising given that the only difference in terms of mRNA sequence between these two sets of constructs is the short 5' UTR downstream of the promoter. Similar results have been obtained for certain mRNAs in yeast (Trcek et al., 2011) and a model was proposed where regulators of mRNA stability are co-transcriptionally loaded onto transcripts and exported into the cytoplasm, which could explain the above observation. An important caveat to the above data is that the control experiment, incubation of cells with DMSO in the absence of actinomycin D,

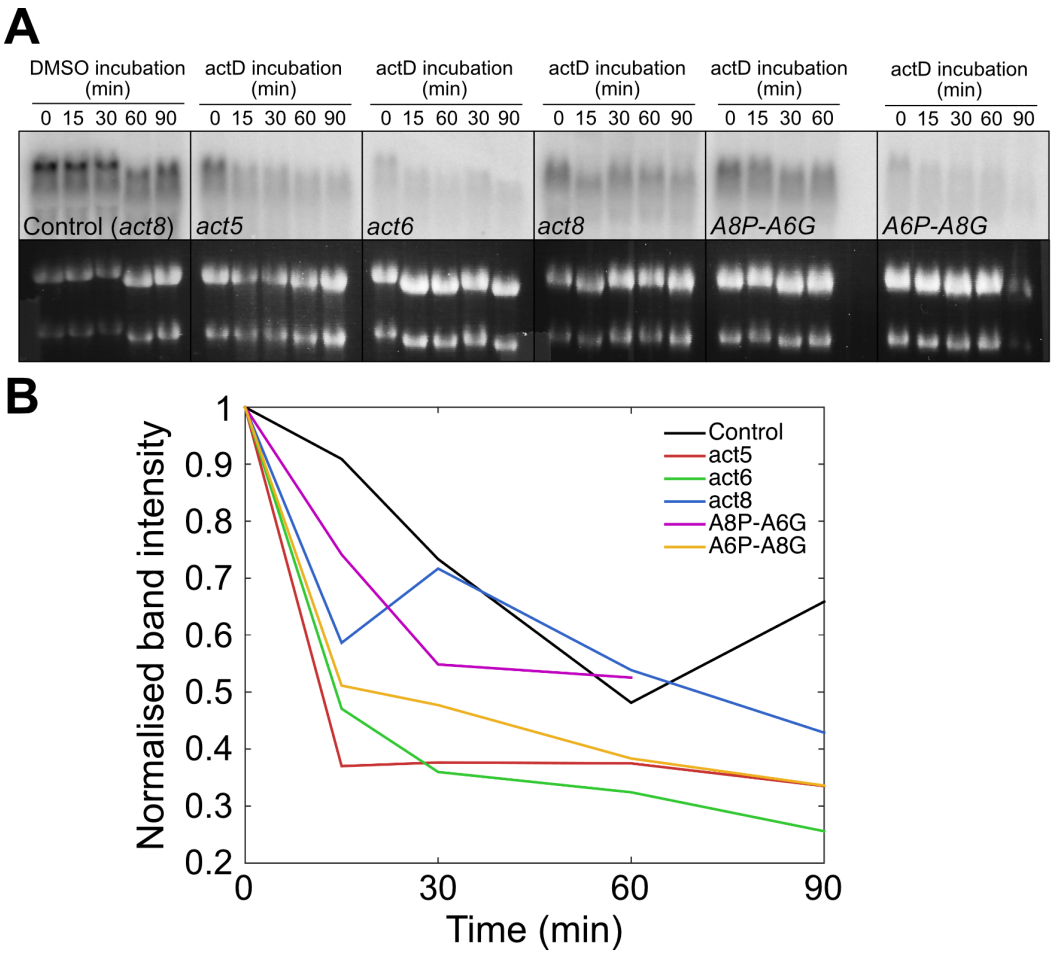

Figure 4.22: Actin mRNA stability determined by actinomycin D assay. (A) Northern blot analysis of actin mRNAs during actinomycin D incubation time course. Cell lines containing MS2-tagged version of each gene were used, and therefore a radioactive MS2 probe enabled the stability of individual actin mRNAs to be determined. Agarose gel photos are shown as loading controls. 90 min sample for A8P-A6G was lost during preparation. (B) Band intensities were quantified by a phosphorimager. Relative decay rates were plotted by normalising band intensity to the beginning of the time course.

also showed mRNA decay and therefore some of the changes seen in actin genes may also be attributed to this.

Neither of the proposed mechanisms tested above appear to explain the reduction in noise differential between nascent transcription and cytoplasmic mRNA populations. It is therefore unclear what is causing this decrease. Alternative mechanisms could include different levels of processing and/or degradation of actin transcripts before they have a chance to leave

the nucleus by RNA surveillance machinery such as the exosome (Chlebowski et al., 2013).

4.11 Generation and characterisation of Actin-mNeonGreen fusion proteins

The reduction in the noise differential between actin genes from nascent mRNA production to cytoplasmic mRNA distributions leads one to question whether this variability in transcription serves any functional purpose. To address this further I generated three cell lines containing Act5, Act6 and Act8 fused to mNeonGreen (Shaner et al., 2013), integrated at the endogenous loci (see section 2.1.8.3). Example fields of view are shown in figure 4.23. As expected, Act8-mNeonGreen cells are the brightest, followed by Act5- and Act6-tagged cell lines, qualitatively matching the expression data discussed previously. All three fusion proteins appear to localise to typical actin-based structures such as macropinocytic cups and crowns (white arrowheads), as well as filopodia or retraction fibres (black arrowhead). These structures, if followed by time-lapse microscopy are dynamic and turnover on the timescale of seconds. From these images there is also very little evidence of aggregate-like foci which might be indicative of dysfunctional protein (aside from the autofluorescence seen in weakly-expressing Act6-mNeonGreen cells). All this suggests that these proteins are probably capable of normal actin function, to some extent at least.

Previous work assessing the functionality of a GFP-Act15 construct in *D. discoideum* showed that while most functional features of actin biology were replicated in the fusion protein, this was dependent on the proportion of labelled monomers in the actin filament (Westphal et al., 1997). To determine whether basic cell biology is affected by expression of actin fusion proteins I measured the growth kinetics of these cell lines compared to control in suspension culture (Fig. 4.24 A). The mean generation time of fusion protein knock-in cell lines are increased compared to control in accordance

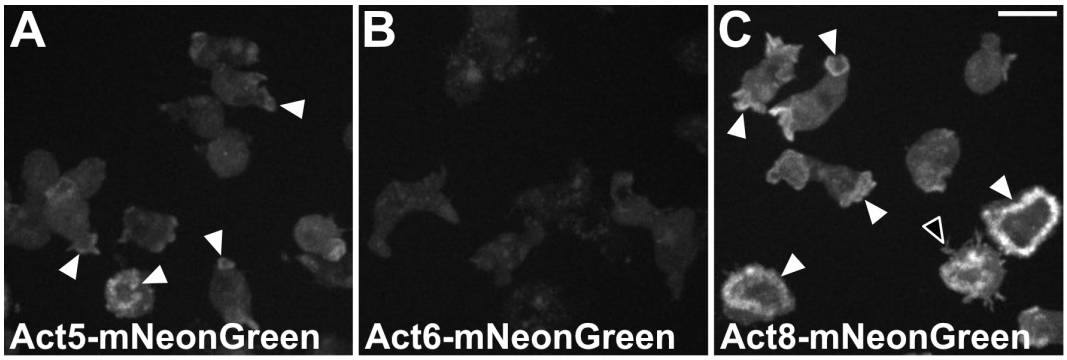

Figure 4.23: Examples of cell lines expressing actin-mNeonGreen fusion proteins. (A) *act5*, (B) *act6* and (C) *act8* fused to mNeonGreen integrated at endogenous loci. Fusion proteins can be seen to localise to typical actin-based structures such as macropinosomes (white arrowheads) and filopodia (black arrowhead). Scale = 10 μm.

with their relative expression levels (Fig. 4.24 B). Act8-mNeonGreen, as the most highly expressed protein, shows the largest growth defect, while Act6-mNeonGreen, which is more weakly expressed, is comparable to the control Ax3 cell line. Therefore, these data are similar to those of Westphal et al. (1997), as high expression of fusion protein will inevitably mean a greater proportion of labelled monomers in microfilaments. One consideration to make when interpreting these data is that growth was measured in suspension culture, which may exacerbate growth defects derived from fusion protein expression. Indeed, anecdotally there was no evidence for slow growth of these cell lines during routine culture on plastic substrates. Furthermore, Westphal et al. (1997) found that only suspension culture led to increased multinucleation in a GFP-Act15-expressing cell line. Therefore, while expression of actin fusion proteins probably affects cytoskeletal function at some level, the impact is likely to be mitigated by culturing cells on surface substrates rather than in suspension. All further experimentation with these cells was done with surface culture.

4.12 Single-cell variability of actin proteins

To determine the distribution of fusion protein expression in a population I used flow cytometry to measure the fluorescence, and thereby protein

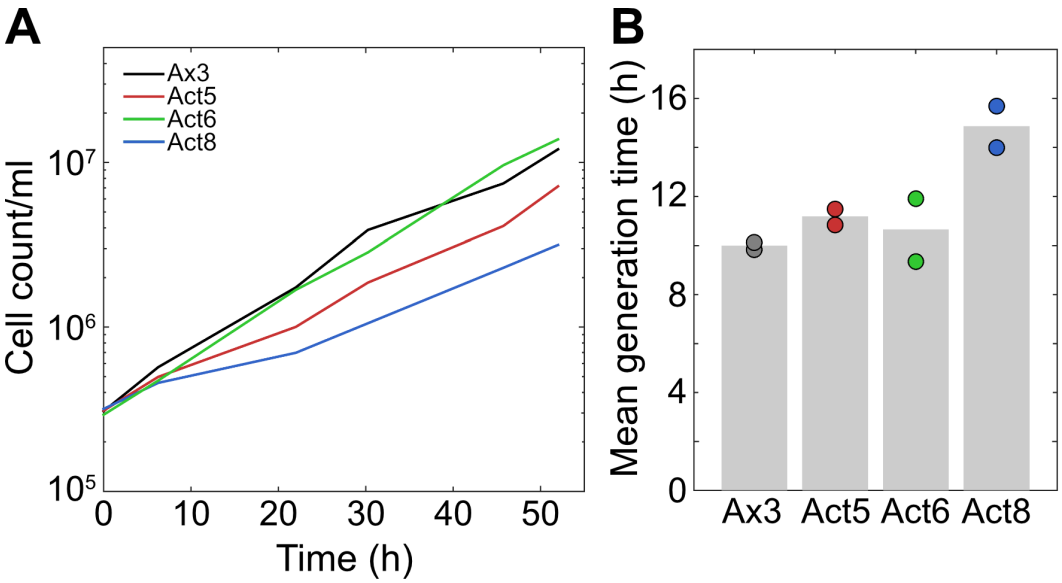

Figure 4.24: Assessment of growth in fusion protein knock-in cell lines. (A) Growth curves of actin-mNeonGreen cell lines compared to wild-type (Ax3). (B) Mean generation time of fusion protein cell lines compared to Ax3 over two experiments. Grey bars are mean, coloured circles are individual experiments.

level, in thousands of single cells. Figure 4.25 A shows a typical example of the fluorescence intensity distributions of actin fusion protein cell populations. The mean intensity values of the different cell lines match the trends described in MS2 and FISH experiments, with Act8-mNeonGreen cells the brightest followed by Act5 and Act6 (Fig. 4.25 A). Therefore, qualitatively, actin protein levels are correlated with mRNA levels. However, the normalised variability of the protein level distributions (C_v^2) are very similar for all actin genes (Fig. 4.25 B). Therefore the large difference in noise between actin genes, initially derived from different transcription dynamics (Figs. 4.5, 4.6) and reduced in cytoplasmic mRNA distributions (Figs. 4.16, 4.17) is effectively absent at the level of the protein (Fig. 4.25). This is quite a striking result given that previous studies of noise in transcription ascribe functional importance to these variable processes, for example in developmental transitions (Losick & Desplan, 2008; Mohammed et al., 2017). However, modelling approaches have also identified translation as the dominant factor in determining both noise and number of proteins in the cell (Thattai

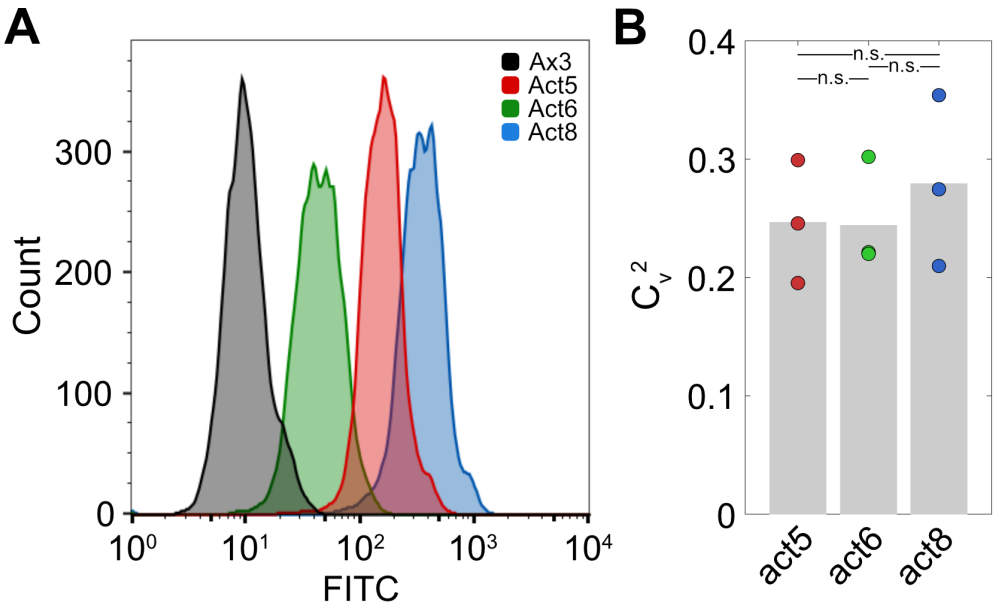

Figure 4.25: Population variability of actin fusion proteins. (A) Representative example of the distributions of single cell fluorescence as captured by flow cytometry for wild-type (Ax3) and actin fusion protein cell lines. (B) Squared coefficient of variation values for fusion protein distributions. Grey bars are mean, coloured circles are individual experiments. Significance determined by paired Student's t-test (Welch's unpaired, unequal variance t-test gave similar results). N.s. = p $>$0.05.

& van Oudenaarden, 2001; Schwanhäusser et al., 2011). Consequently, for a given gene, fluctuations in transcriptional activity could be either amplified or dampened depending on fundamental downstream processes such as mRNA degradation, translation and protein degradation. This could depend on the function of the protein product; for a housekeeping protein such as actin it may be beneficial for the cell to maintain transcriptional noise at a minimum.

4.13 Summary

In the above chapter I generated cell lines in which individual actin genes are specifically targeted with reporter constructs in order to assess dynamic behaviours of gene expression. Principally, I made use of the MS2 system in which a series of encoded RNA stem-loop structures are integrated into a gene of interest to enable the live imaging of the transcriptional activity

of that gene (Fig. 4.1). I targeted three new actin genes with this system and made use of a fourth cell line previously constructed in the lab (Fig. 4.2). I showed that actin genes display at least two distinct mechanisms of transcription (Fig. 4.4) with large transcriptional bursts following periods of inactivity (*act5*, *act6*) contrasted by near-constitutive activity (*act8*). I then characterised these dynamics more quantitatively, showing that the distribution of transcriptional activity is highly variable in 'bursty' genes and less so in *act8* (Figs. 4.5, 4.6).

Next, I investigated the determinants of these differing transcriptional dynamics. Having seen in Chapter 3 that actin genes have divergent promoter regions which retain differing combinations of putative regulatory elements, I generated cell lines in which the promoters of *act6* and *act8* are integrated in place of the endogenous promoter of the other gene. Transcription dynamics in these cell lines are remarkably similar to those exhibited by the promoters situated at the endogenous loci (Fig. 4.8). Furthermore, the shape and relative variance of the spot intensity distributions as well as cooccurrence matrices are highly similar to those of the promoter at the endogenous locus (Figs. 4.9, 4.10). Therefore, promoter sequence seems to be the dominant factor in controlling transcriptional dynamics in *Dictyostelium* actin genes. Known effects of certain promoter sequence features such as poly(dT) tracts, TATA boxes and number of putative regulatory motifs all correlate with the observed dynamic behaviours (Fig. 3.7).

Beyond this, I also investigated how other factors can influence transcription dynamics. Transcriptional activity is weakly correlated with both cell size and speed for all genes, explaining some of the variability in the population (Figs. 4.11, 4.13). Most strikingly, while *act5* and *act6* transcriptional activities show a clear negative correlation with population density, *act8* gene activity appears to be unaffected by density changes (Fig. 4.12). Despite the relatively small sample size and a potential confounding factor in density-dependent changes to speed (Fig. 4.14), this could demonstrate

different roles for actin genes in the cell, with *act8* required for basal protein production, and *act5* and *act6* responsive to environmental cues. The different transcriptional behaviours of the genes are further suggestive of this.

Having identified distinct differences in the variability of transcription dynamics, I next looked at whether these differences propagate through the subsequent processing steps in protein production beyond transcription. Using single molecule RNA FISH (Fig. 4.15) I measured the cytoplasmic mRNA distributions of cell populations for different actin genes (Fig. 4.16). Interestingly, the difference in variability, or noise, between the actin genes was reduced (Fig. 4.17). As for nascent transcription, cytoplasmic mRNA distributions of promoter-switched cell lines broadly matched those of the gene from which the promoter derives (Fig. 4.18). However, aggregates or granule-like structures found in one promoter-switched cell line, which were not present in FISH experiments of either of the 'parent' genes, suggest some interaction between 5' and 3' UTRs in these transcripts (Fig. 4.19). I also checked the relationship between cell size and gene activity using this method and showed a stronger correlation for all genes than transcriptional dynamics alone (Fig. 4.20). The slope of the linear regression between cell size and mRNA count was significantly higher for *act8* compared to *act5* which may tie in with a role as the producer of basal levels of actin protein.

To examine the reduction in the noise differential levels between actin genes from nascent mRNA production to steady-state distributions I correlated counts of nuclear and cytoplasmic mRNAs. While there was little difference in the relationship between these counts for the highly expressed genes, *act1* showed a significantly increased gradient of the linear regression (Fig. 4.21), potentially suggestive of differences in nuclear export. However, this does not explain the effect on noise as described. I then measured the decay rates of the mRNAs by an actinomycin D assay. The differences in mRNA stability between actin genes actually appear to be opposite to

what might be expected if it were to explain the noise differential reduction. Therefore, there are likely to be other mechanisms contributing to this effect.

Finally, to understand whether the previously described differences in transcription dynamics and mRNA distributions are visible at the protein level I generated fusion proteins of actin genes and mNeonGreen. The fusion proteins localise to actin-based structures (Fig. 4.23) and while the more highly-expressed genes show a growth defect in suspension culture, this is probably minimised when culturing on a plastic surface (Fig. 4.24). The variability of protein level distributions as assessed by flow cytometry were identical for all actin genes (Fig. 4.25). Therefore a noise differential introduced during transcription appears to be lost during the many subsequent steps in gene expression.

Overall, there are at least two potentially important points raised by these data, particularly in the context of an expanded gene family. Firstly, having multiple copies of essentially the same gene seems to facilitate broad responsiveness of the system to multiple stimuli, as demonstrated by the different transcription dynamics. This could be particularly important in a highly compact genome (Eichinger et al., 2005) in which there are currently no reported long-range control elements. However, high levels of noise introduced by transcriptional bursts could be problematic for the cytoskeleton as an essential housekeeping component of the cell. Therefore, by reducing the difference in noise levels across the gene family post-transcriptionally, the cell can engineer robustness into an otherwise noisy system.

Chapter 5

Phenotypic characterisation of actin gene knockout cell lines

5.1 Introduction

Other examples of large multigene families encoding multiple copies of the same gene are often proteins which are typically required at high concentrations within the cell. These include histones (Marzluff et al., 2002) and ribosomal RNA genes (Eickbush & Eickbush, 2007). Indeed, as ribosomal RNA is thought to make up as much as 80% of total RNA in a yeast cell (Warner, 1999), having multiple genes encoding the same RNA or protein would appear to be a natural consequence of such high transcriptional demand. In chapter 4 I showed that extensive duplication of actin genes in *D. discoideum* has probably enabled increased responsiveness of the gene family to external stimuli, and therefore may serve a regulatory function. However, actin makes up around 10% of all protein in *Dictyostelium* cells (Uyemura et al., 1978) and a further role for a large gene family as a mechanism with which to cope with such demand is not unreasonable. In the following chapter I have explored this idea using cell lines with multiple actin genes either deleted or disrupted to reduce the level of the protein within the cell. I then characterised the effects of actin depletion on various aspects of cell biology in order to determine which of the many actin-dependent processes are most sensitive to loss of the protein.

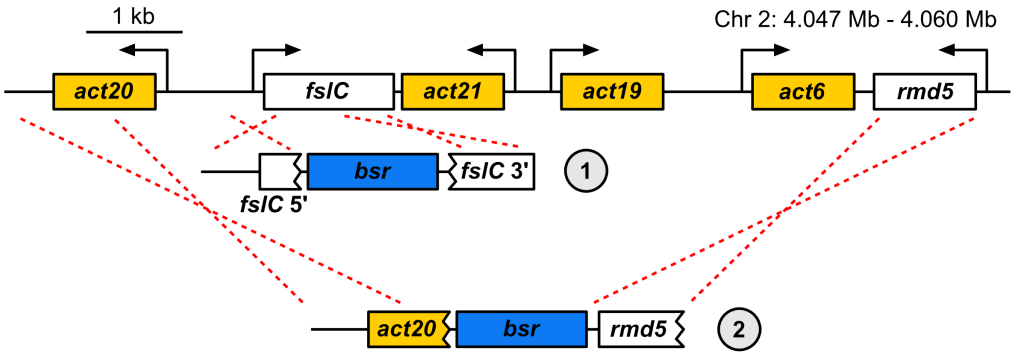

Figure 5.1: Deletion of four actin genes by a single targeting event. Four actin genes (yellow) are situated within 10 kb of each other on chromosome 2. To delete all four genes in one targeting event I generated a construct (2) to recombine out the actin genes, replaced with *bsr* to enable selection. Given that this process also deletes *fslC* I generated another construct (1) as a control for effects of the actin deletion. Detailed maps of the constructs (1) and (2) are shown in figure 2.5.

5.2 Generation of actin gene knockout cell lines

In order to reduce actin protein levels in *D. discoideum* cells I generated several knockout cell lines. Given that there are 17 genes encoding the major actin subunit in *D. discoideum* I assumed that several of these would need to be knocked out in order to significantly reduce mRNA and protein levels. Therefore, to disrupt multiple actin genes in the most efficient manner I took advantage of the fact that four actin genes are situated adjacent to one another on chromosome 2 (Fig. 3.2). As the schematic in figure 5.1 shows (see section 2.1.8.4 for more details), I generated a deletion construct to remove all four genes in a single targeting event. Unfortunately, use of this construct would also mean deletion of an unrelated gene situated between *act20* and *act21*, *fslC*. Therefore, I also generated a construct to disrupt this gene alone, as a control for any effects in the 4x actin gene knockout. *fslC* is a frizzled- and smoothened-like 7-transmembrane domain G-protein coupled receptor, part of a family of 25 similar proteins in Dictyostelium (Prabhu & Eichinger, 2006). The only reported mutant of one of these proteins showed no developmental phenotype (Sawai et al., 2007), and little else is known about the role(s) of these proteins.

Given that the cell lines used for monitoring gene expression dynamics in Chapter 4 also disrupt the coding sequence (Fig. 4.1) and are therefore effectively gene knockouts, I combined this 4x gene knockout with these cell lines to generate a suite of knockout cell lines. In particular I focussed on an *act8* knockout (*act8* targeted with PP7 repeats, stem-loop-forming DNA sequences analogous to MS2, in a wild-type Ax3 background as opposed to H2B-mCherry KI background as described in Chapter 4), the 4x knockout and a cell line with six disrupted actin genes (genotype: Ax3 background, *act5*-PP7, *act8*-MS2, 4x actin KO).

5.3 Reduced mRNA content in actin gene knockouts

One potential hypothesis to explain the size of the actin gene family is that having multiple genes could enable compensation for each other in terms of gene expression, ensuring robust expression the protein. If actin genes can compensate for each other's expression to maintain optimal protein levels in the cell, it would suggest that the large gene family is not required to facilitate the high levels of expression in *D. discoideum* cells (Uyemura et al., 1978; Margolskee & Lodish, 1980). To test whether gene disruption actually reduces the level of actin within cells, and therefore whether gene loss can be compensated for, I measured actin mRNA levels by northern blot in wild-type and knockout cell lines. While the 4x *act* knockout and the Δ*fslC* control show little difference to wild-type, disruption of six actin genes, including the highly expressed *act5* and *act8*, slightly reduces actin mRNA levels (Fig. 5.2). Unfortunately this experiment was only done once and band intensities were not measured by phosphorimager meaning differences between cell lines could not be quantified. However, there is clearly a reduction in band intensity, and more specifically, an almost complete loss of intensity in the lower band of the actin mRNA doublet in 6x *act* knockout cells (Fig. 5.2). This correlates with the fact that *act8* is one of the genes predicted to have

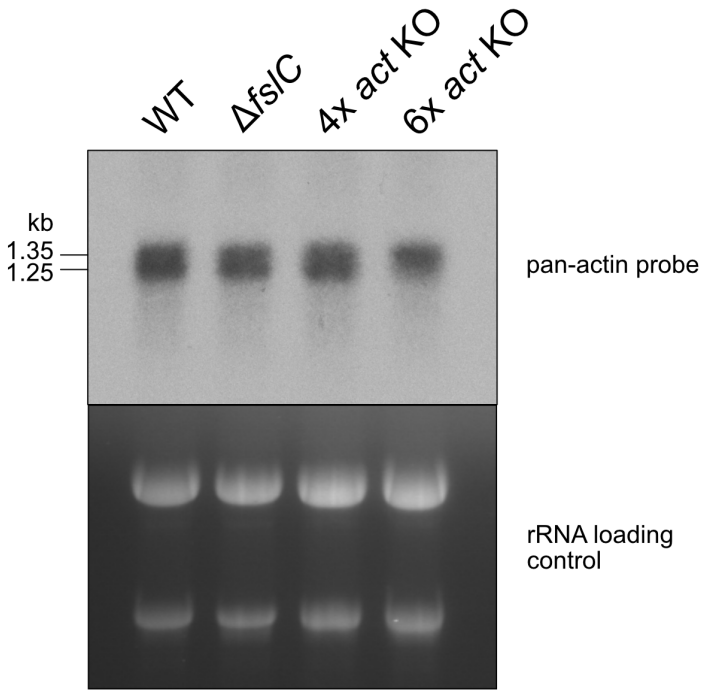

Figure 5.2: Actin gene expression in knockout cell lines. Total actin mRNA levels were assessed by northern blot in actin knockout cell lines. A pool of pan-actin probes was generated by PCR using primers which hybridise to all actin genes. The agarose gel photo with strong rRNA staining is shown as a loading control.

a shorter 1.25 kb transcript (Romans et al., 1985). Therefore, actin gene disruption does reduce overall actin mRNA levels suggesting that the maintenance of a large family of actin genes is unlikely to be required to enable compensation between gene copies. However, this is largely based on qualitative analysis and a single experiment and therefore should be repeated to be more conclusive about this.

5.4 Growth kinetics of actin gene knockouts

Having seen changes to the actin cytoskeleton can affect growth rate in *Dictyostelium* (Fig. 4.24) and, more generally, knowing that actin is important for cytokinesis among other processes (Pelham & Chang, 2002) I measured the growth rates of actin gene knockout cell lines. As figure 5.3 shows, knockouts of *fslC*, *act8*, and the four actin genes in figure 5.1 show no difference in growth rate to the Ax3 control. While not statistically signif-

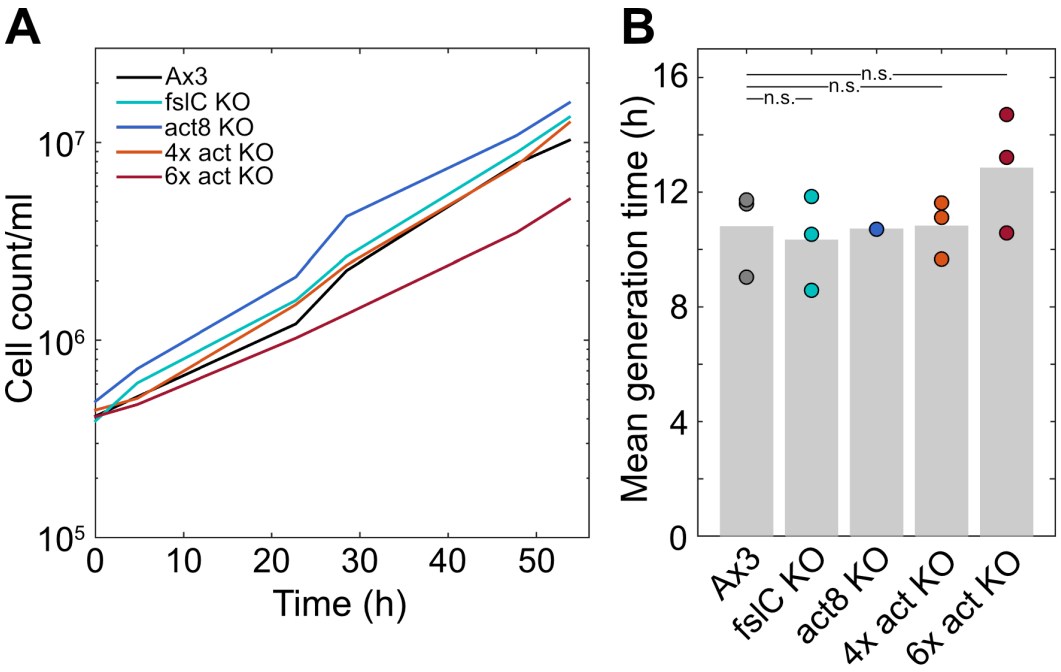

Figure 5.3: Growth rates of actin gene knockout cell lines. (A) A representative example of actin gene knockout growth curves. (B) Mean generation times of knockout cell lines relative to wild-type. Grey bars are mean, coloured circles are individual experiments. Significance was determined using paired Student's t-test.

icant (p = 0.063, paired Student's t-test) the 6x gene knockout consistently showed slower growth than the wild-type in all experimental repeats (Fig. 5.3). This suggests that there is probably a weak effect of disrupting six actin genes (as well as *fslC*) on growth. However, as noted in section 4.11 this effect may only be visible in this assay because of the difficulties of growth in suspension culture. Regardless, it seems that a reduction in actin mRNA levels as shown in figure 5.2 causes a mild defect in *D. discoideum* growth in suspension culture. Perhaps unsurprisingly, growth defects have been demonstrated as a result of actin knockdown in other systems (Bunnell et al., 2011) but the effects here are minimal in comparison, probably because of the relatively small knockdown of protein.

5.5 Cell motility in actin gene knockouts

To further characterise the effects of the various actin gene disruptions on *Dictyostelium* biology I monitored cell motility in the knockout cell lines. To

do this I harvested cells from pre-clearing bacterial agar plates (where the bacterial food supply is still in abundance and none of the cells are starving) as bacterially-grown cell motility is known to be faster than axenically grown cells (Varnum et al., 1986). Using an automated software plug-in in ImageJ I tracked individual cells over 1-2 hours and calculated the average instantaneous speed of cells over 4-8 fields of view. Initially, cells were imaged immediately after plating and adherence to plastic 6-well dishes (having been washed several times in KK2; 'normal protocol' in section 2.5.2) (Fig. 5.4 A, C). In the example shown (Fig. 5.4 C), wild-type Ax3 cells showed an increase in motility 10 minutes after plating, which then plateaus after about 70 minutes as cells fully adjust to their surroundings. Strikingly, both a single *act8* knockout and the 6x *act* knockout show both increased speed and a more rapid resumption of motility relative to the wild-type (Fig. 5.4 C). This is contrary to motility phenotypes of most other actin cytoskeleton mutants (Noegel & Schleicher, 2000). This effect was reproducible over multiple experiments and several different actin knockout cell lines. However, the speed of the cell lines appeared to be dependent on the number of bacteria present in the imaging dish (which varied significantly between experiments), having escaped removal during the washing protocol (see Fig. 5.4 A). Therefore, I developed a new protocol (see 'stringent protocol' in section 2.5.2) involving many more washes to attempt to reduce this background level of bacteria. Figure 5.4 B shows an example FOV after these extra wash steps, with far fewer bacteria per unit area compared to figure 5.4 A. Upon repeating the tracking and motility analysis it became clear that under these conditions there was no difference in the speed of cells (Fig. 5.4 D; cells imaged 2 h post-plating). The motility phenotype is therefore dependent on the presence of bacteria. This could implicate other processes such as phagocytosis but further experiments would be required to test this explicitly.

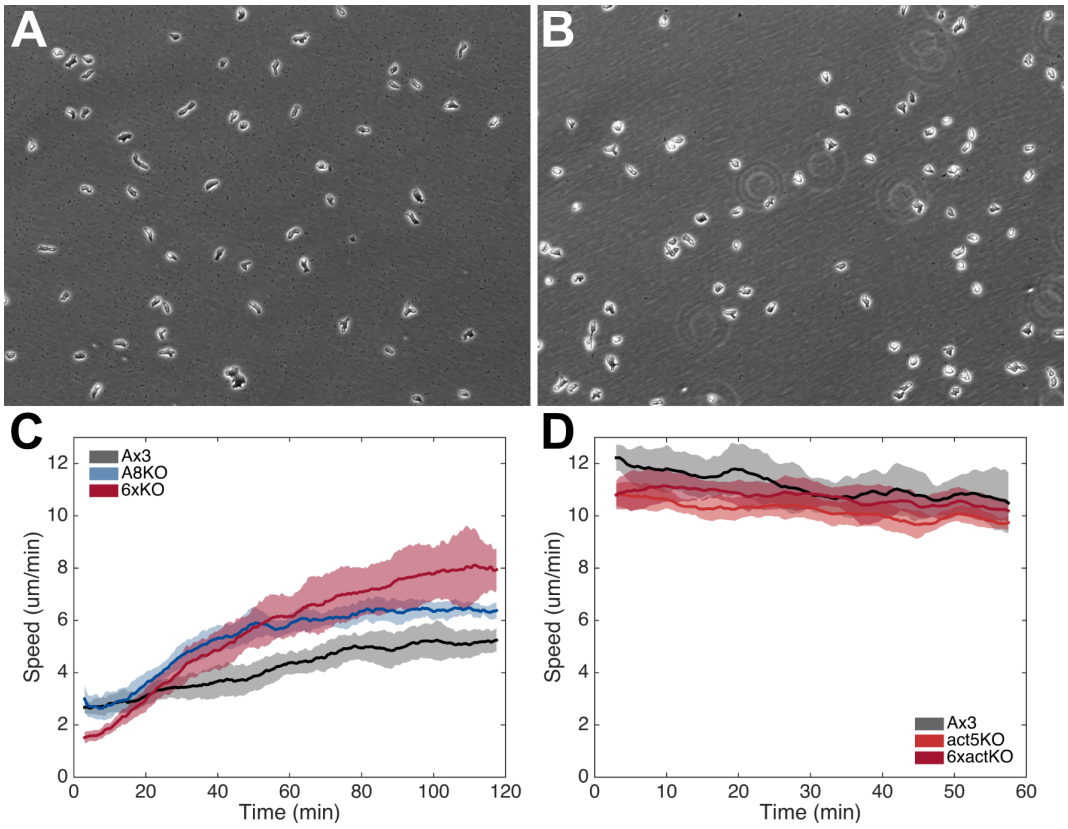

Figure 5.4: Actin gene disruption effects on cell motility. (A) Example field of view of wild-type cells prepared using initial protocol. (B) Example field of view of cells prepared with more stringent washing protocol. (C) Average speed of cells across six fields of view for each cell line. Shaded areas show one standard deviation. Imaging was initiated immediately after plating cells. (D) Same as (C) but cells prepared with stringent washing protocol and imaged 2 h after plating.

5.6 Development of actin gene knockouts

Given that the actin cytoskeleton is an important component of the cell adhesion machinery, and that *Dictyostelium* development is strongly dependent on cell-cell adhesion, I tested whether knocking out actin genes affected developmental progression. Figure 5.5 shows representative images of a single development experiment comparing Ax3 and 6x *act* knockout cells at 7 h of development. At this time point both cell lines have started streaming within aggregation territories and mounds are beginning to form at the centre. End point assessment of development at 24 h similarly showed both cell lines capable of forming mature fruiting bodies by this time. Therefore

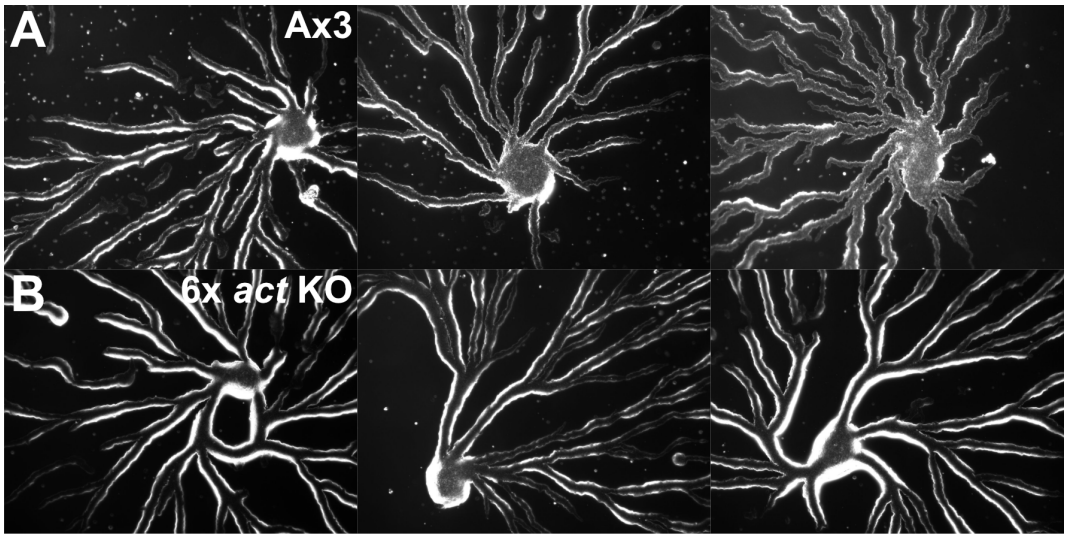

Figure 5.5: A comparison of development in wild-type and actin knockout cells. Wild-type (Ax3) and cells with six actin genes knocked out were developed on KK2 agar for 7 h before imaging. Representative images of developmental structures are shown. No developmental delay was observed in actin knockout cells compared to wild-type.

it would appear that disruption of 6 out of 17 actin genes has no effect on developmental progression, despite the strong upregulation of actin genes during early development as indicated by RNAseq (Fig. 3.10). This could be due to the relatively mild knockdown of actin expression (Fig. 5.2) and therefore further disruption of the remaining actin genes may result in a developmental phenotype.

5.7 Summary

In this chapter I have explored the effects of actin gene knockouts on *Dictyostelium* biology. To achieve this I exploited the fact that four actin genes are tightly clustered within 10 kb on chromosome 2 and generated cell lines with all four deleted from a single targeting event (Fig. 5.1). I combined this deletion construct with other cell lines with MS2- or PP7-tagged genes (in which the coding sequences are disrupted) to knock out up to six actin genes in *D. discoideum* cells. Firstly, to test whether knocking out actin genes actually reduces the overall level of actin expression, and therefore whether multiple genes are required for the high protein levels in the cell,

I measured total actin mRNA in knockout cell lines by Northern blot (Fig. 5.2). Cells with six actin genes disrupted showed reduced levels of actin mRNA, suggesting a requirement for multiple genes to enable high levels of actin gene expression. I then characterised the effects of this disruption of six actin genes (as well as the *fslC* gene) on actin-dependent processes. In cells which have only *act8* or four relatively-weakly expressed genes disrupted there appears to be no discernible defect in growth rate in suspension culture. However, in a cell with six actin genes, including *act8*, knocked out there is a slight, but reproducible growth defect (Fig. 5.3). Using an assay to measure random cell motility I found that actin knockouts move faster than wild-type (Fig. 5.4 A, C). Subsequent optimisation showed that this effect was dependent on the presence of bacteria, suggesting that the motility phenotype might influenced by defects in other parts of the cell (Fig. 5.4 B, D). Finally, there appears to be no developmental phenotype in the knockout cell lines (Fig. 5.5). Overall, these data show that there is likely to be a requirement for a large actin gene family to enable high levels of protein production, as well as the regulatory role outlined in Chapters 3 and 4. While the phenotypes of actin gene knockout cells are relatively weak, this is probably due to the comparatively mild knockdown of actin by disruption of only 6/17 genes. Further cumulative disruption of family members could exacerbate the phenotypes and demonstrate this more clearly.

Chapter 6

Discussion

6.1 Summary of results

This work has demonstrated a potential role for differential transcriptional regulation in the expansion of a highly expressed gene family encoding the same protein isoform in multiple genes. The actin gene family of *Dictyostelium discoideum* contains 17 genes encoding an identical protein. Other dictyostelids were also found to have large actin gene families containing multiple genes for a single isoform, similar to other amoebas, implying that this is a general feature of amoeboid genomes. Despite having identical amino acid sequences, differential codon usage may facilitate regulation at the level of translation.

A detailed analysis of the 5' and 3' flanking regions of the genes encoding identical proteins highlights strong conservation of individual sequence motifs amid significant divergence overall. Different combinations of promoter elements may permit variable transcriptional regulation, while downstream putative regulatory elements show two distinct patterns which are suggestive of differential 3' end processing. Only subtle differences in developmental expression pattern across the gene family were revealed by RNAseq, yet these differential patterns correlated with the presence of different combinations of putative regulatory elements in the gene promoters. Single cell RNAseq showed highly correlated expression of most actin genes in single cells, which argues against the idea that these genes are

able to buffer each other's expression.

Using the MS2 system to assess the transcriptional behaviour of a number of actin genes showed that these genes are transcribed with different dynamics in growth conditions. Noisy gene expression of some actins, derived from transcriptional bursts, is contrasted by more constitutive behaviour of another gene. Further analysis showed that this constitutive behaviour, rather than following a Poisson distribution of transcription initiation events, is controlled by transcriptional bursting but the time between bursts is so short as to appear constitutively active. Promoter sequences appear to control the majority of the transcriptional dynamics, as a promoter switching experiment almost completely reverses the dynamic behaviour of the endogenous locus. Cell size explains only a small amount of the variability in gene expression for all three genes. However, bursty genes may be responsive to environmental factors such as population density, with the constitutive *act8* gene seemingly unresponsive.

Comparing the noise levels between genes from nascent transcription dynamics to cytoplasmic mRNA distributions shows that the large difference in noise between bursty and constitutive genes is reduced at this later stage of gene expression. Promoter-switched genes are similarly distributed at this stage. This reduction in noise does not appear to be due to differences in mRNA stability as determined by an actinomycin D assay, or transcript retention in the nucleus. Measurement of variability between cells at the protein level by generation of actin-mNeonGreen fusion proteins reveals complete loss of the noise differential between actin genes. This suggests that having multiple actin genes enables environmental responsiveness by integrating different signals during transcription, but also enables robustness as the noise introduced by transcription is filtered out during mRNA processing and protein production.

Dictyostelium cells with several actin genes knocked out showed reduced mRNA levels by northern blot again suggesting that these genes

are limited in their ability to compensate or buffer each other's expression. Knockout phenotypes were only clearly displayed in cells with six actin genes disrupted and even then the effects were mild. These cells showed slower growth and faster motility in the presence of bacterial prey (but not in their absence, potentially suggestive of a phagocytosis defect) but no developmental defect. These data suggest that multiple actin genes are also required to produce the high protein levels required for normal cell physiology on top of the regulatory role described above.

6.2 Actin gene family evolution in Amoebozoa

One striking observation from this work which tallies with previously published data (Lahr et al., 2011; Gunning et al., 2015) is that almost all genomes of members of the Amoebozoa contain actin gene families at least partially comprised of multiple genes encoding identical an isoform. Of the nine dictyostelid species surveyed here, eight match this description (Table 3.1), while two non-dictyostelid Amoebozoa species, *Entamoeba histolytica* and *Arcella* sp. exhibit the same phenomenon (Lahr et al., 2011; Gunning et al., 2015). This appears to be exclusive to amoebas, as other protists and members of other kingdoms do not have the same actin gene family organisation (Gunning et al., 2015). The evolutionary distance between *Dictyostelium* species is considerable (*D. discoideum* and *D. purpureum*, both group 4 dictyostelids, are separated by around 400 million years (Parikh et al., 2010)), while gene family expansion was shown to have occurred independently between *D. discoideum* and *D. fasciculatum* (Joseph et al., 2008). The gene families in these organisms appear to follow a birth-and-death model with purifying selection given the high number of synonymous substitutions and presence of many pseudogenes (Nei & Rooney, 2005). These observations strongly suggest that this type of actin gene family architecture has been selected for during evolution. A more expansive survey of published amoeboid genomes would lend further strength to this idea.

This leads one to ask what features of amoeboid biology require such

an arrangement? Certainly, these cells appear to need a lot of the protein, with actin making up around 8-10% of all protein in *D. discoideum* (Uyemura et al., 1978), which is at the higher end of estimates of actin content in other cell types (Hanson & Huxley, 1957; Bray & Thomas, 1975; Rubin et al., 1978; Stark et al., 1982; Varani et al., 1983; Gowing et al., 1984). Furthermore, *Dictyostelium* cells (and other amoebas (Zaki et al., 2006)) are highly motile, feed by phagocytosis and macropinocytosis and acquire strong cell adhesive properties during development, all processes which depend on actin for proper functioning. However, leukocytes utilise many of the processes described above (Friedl et al., 2001; Artemenko et al., 2014; Skoge et al., 2016), and have similar levels of actin protein (Stark et al., 1982; Varani et al., 1983), and yet the six separately encoded isoforms of actin (in mammals) are sufficient in these cells. What could account for the differences in actin gene families between these two similar cell types? One explanation could be the differences in variability of the local environment experienced by these cells. *Dictyostelium* cells are typically isolated from forest floor soil samples (Raper, 1951) and are presumably subject to considerably more diverse environmental conditions than the homeostatic context of a mammalian body, hence the different life cycle responses to starvation. As will be discussed further later, having multiple diverse promoters driving the expression of the same coding sequence could allow the integration of signals from different stimuli in this variable environment, particularly in the context of a highly compact genome (Eichinger et al., 2005). An alternative explanation is simply that the growth rate of *Dictyostelium* cells is much higher than those of animals and therefore increased actin production is required to accommodate this.

How do these data fit into to the arguments of Gunning et al. (2015) concerning the evolution of multiple mechanisms to achieve functional specificity of actin and its binding proteins in the different kingdoms of life? Gunning et al. (2015) noted that while plants typically have a greater num-

ber of isoforms of both actin and its regulators, yeast and animals utilise tropomyosin to generate functional diversity in actin structures. In protists, which also lack tropomyosin, the number of actin and actin-binding protein isoforms appear to be fewer than plants (but more than animals and yeast), so how then is such functional diversity achieved? One, perhaps speculative, hypothesis is that differential codon usage might be involved. In this work, I showed that actin genes with more 'optimal' codon usage are weakly expressed in growth conditions and strongly induced in development, while strongly expressed genes in vegetative growth are those with less optimal codon usage (Figs. 3.3,3.10). Differential codon usage has been shown to affect multiple facets of gene expression including mRNA stability (Presnyak et al., 2015), translation efficiency (Sørensen et al., 1989; Yu et al., 2015; Buhr et al., 2016) and subsequent protein folding and structure (Zhou et al., 2013; Kim et al., 2015; Fu et al., 2016) in a range of model organisms. On top of this, the use of rare or non-optimal codons has been shown to be under selection (Clarke & Clark, 2010; Pechmann & Frydman, 2013). If the differences in actin gene codon usage are both selected for and functional, what might this mean for the organism? In theory, structural differences introduced by altered codon usage and therefore translation rates (as demonstrated previously (Zhou et al., 2013; Kim et al., 2015; Fu et al., 2016)), could result in changes to binding sites on the actin monomer, enabling diversification of binding partners between otherwise identical coding sequences. The fact that the majority of the weakly expressed, highly adapted cluster of genes in figure 3.3 are those which are expressed transiently in early development (Fig. 3.10) could be further indication of a unique functional requirement at this stage. Alternatively, as the mRNAs from these genes are predicted to be translated faster, their transient expression here could simply represent a requirement for rapid protein production in early development. Given most of this is conjecture, specific experimental evidence would be required, firstly that different actin genes are transcribed at dif-

ferent rates, and secondly that, if true, this has any effect on the protein molecule and its structural binding motifs, in order to be conclusive either way.

6.3 Promoter-derived transcriptional dynamics of actin genes

By using live imaging techniques to monitor dynamic behaviour in real-time I showed that individual actin genes are transcribed by different mechanisms (Fig. 4.4). Despite the coding sequence of these genes apparently being under strong selective pressure, their regulatory control has not been strictly maintained, with the extent of variability or 'noise' in transcriptional activity differing between genes (Fig. 4.6), mediated by the size and frequency of transcriptional bursts. These data suggest that actin gene expression is regulated by inputs from multiple stimuli.

To explore the determinants of such differential transcriptional behaviour I switched the promoters of two actin genes and repeated the analysis in these cells. Transcription dynamics of genes expressed from the same genomic locus but with a different actin promoter reflect those of the promoter at its endogenous locus (Figs. 4.8,4.9). Therefore, at least in this system, transcription dynamics are determined almost exclusively by the core promoter and upstream regulatory elements, and occur independently of genomic locus. Previous studies have shown a similar phenomenon of variable dynamics with alternative promoters by live imaging with an endogenous and CMV promoter (Yunger et al., 2010) and inferred by RNA FISH experiments in fixed cells using several unrelated promoters (Hocine et al., 2015). However, in both cases only a single locus was examined, while non-natural promoters were used in each case. Therefore, this study provides clearer evidence of sequence-specific transcriptional bursting at multiple genomic loci. This is in contrast to other studies using random integration of reporter genes which identified position effects on both burst

size and frequency at variant genomic loci (Skupsky et al., 2010; Batenchuk et al., 2011; Dar et al., 2012). However, these studies all used (destabilised) GFP fluorescence intensity as a readout and subsequently inferred dynamic behaviour from protein level fluctuations or distributions. On average, protein half-lives are considerably longer than mRNAs (Schwanhäusser et al., 2011) and neither protein nor mRNA distributions are necessarily a good readout for nascent transcription dynamics, as demonstrated in this work. While promoter sequence appears to control most of the variability in transcription, comparison of temporal fluctuations by analysing cooccurrence matrices (Fig. 4.10) shows some differences which could be explained by local chromatin environments or other position effects. Indeed, transcriptional burst size of *act5* was altered in a Set1 (histone H3K4 methylase) knockout in *Dictyostelium* (Muramoto et al., 2010) but these cells have other phenotypic defects and so the effect on transcription could be secondary to these.

While the result of the promoter switching is quite clear, the experiment itself is fairly crude. Further experimentation could lead to more precisely uncovering the specific determinants of the different dynamic behaviours. Promoter sequences are often the basis for recruitment of numerous factors involved in the initiation and regulation of transcription. For example, the core promoter is a region around 30 bp either side of the transcription start site (TSS) and typically contains conserved sequence elements which serve to recruit the general transcription machinery (Thomas & Chiang, 2006). The TATA box, which is the most well-studied promoter sequence element, is specifically bound by the TATA-binding protein (TBP), often as part of TFIID, a general transcription factor (GTF) complex. The TATA box has been shown to be associated with variable transcription (Raser & O'Shea, 2004; Blake et al., 2006; Tirosh & Barkai, 2008; Hornung et al., 2012), although a recent genome-wide analysis in *Dictyostelium* found that this was only visible after correction for gene length (Antolović et al., 2017). Studies have

demonstrated a role for consensus TATA box sequences in rapid reinitiation of transcription, for example (Yean & Gralla, 1997). These high affinity sequences are predicted to influence noise by affecting the specific complexes into which TBP is bound at the promoter, the high turnover rates of which influence transcriptional noise (Ravarani et al., 2016). Indeed, TATA box mutations at the *act5* promoter in *Dictyostelium* prevented the attainment of the very highest initiation rates, although transcription was still highly variable in these cells (Corrigan et al., 2016). Therefore, the TATA box is likely important for generating transcriptional noise but other sequence-based factors are also likely to be involved. These include: the number of transcription factor binding sites upstream of the core promoter (Raj et al., 2006; Sharon et al., 2014; Wolf et al., 2015), where competition for binding within a clustered regions of sites may generate variability in initiation rates; and the nucleosome occupancy of the promoter (Bai et al., 2010; Brown et al., 2013; Small et al., 2014) which in turn is associated with poly(dT:dA) tracts (Struhl, 1985) but also chromatin modifiers and remodellers which can be recruited by transcription factors and RNA polymerase II alike (Cairns, 2009). The combination of these features in the promoters surveyed here correlates with their proposed roles in noise regulation, with noisy genes containing TATA boxes, many putative binding sites and less poly(dT:dA) tracts and vice versa for quiet genes. Specific manipulation of individual sequence features could more precisely determine their influence of transcriptional bursting.

Why would an organism regulate genes encoding identical proteins differently? Some evidence exists to show cell type-specific expression of otherwise identical histone H4 genes in humans (Holmes et al., 2005). However, *Dictyostelium* are single-celled organisms and this differential regulation occurs even in vegetative growth conditions. An alternative explanation may lie in the requirement for robust responses to the highly variable environment in which microorganisms typically find themselves. Various studies have shown that environmentally-responsive genes such as those involved

in stress-responses are more likely to utilise noisy gene expression, while genes involved in growth and metabolism are less noisy (López-Maury et al., 2008; Balázsi et al., 2011). In particular, TATA box-mediated variability was shown to be directly responsible for mediating population survival in harsh conditions when positioned upstream of an antibiotic resistance gene (Blake et al., 2006). Thus, noise in gene expression can be beneficial for responding to heterogeneous environmental conditions. Here, I showed that certain sources of heterogeneity, such as cell size, contribute minimally to variability in nascent transcription (r^2 = 0.14, 0.08, 0.06 for *act5*, *act6* and *act8* respectively), meaning most of the variability at this level is likely derived from other inputs. Furthermore, I showed some evidence to suggest that only noisy actin genes (*act5* and *act6*) are responsive to population density (Fig. 4.12) although more data are needed to test this explicitly. Overall this could suggest that *Dictyostelium* cells contains some actin genes which are responsive to environmental stimuli, while others are less responsive and perhaps simply required for basal protein production.

This gene family architecture may represent some form of subfunctionalisation or division of labour with respect to regulatory control, with noisy transcription enabled by relaxed constraints on gene expression in the ancestral organism (Lehner, 2010). Indeed, such responsiveness is likely to have evolved over time given that less variable promoters appear to be the default configuration, and that noise is proposed to evolve as a response to variable environments (Wolf et al., 2015). Duplication within the gene family may have further facilitated this as duplicate genes have been shown to be more noisy in comparison to singletons (Wapinski et al., 2007; Dong et al., 2011). This promoter evolution could be an alternative strategy to the use of multiple enhancers to confer robust gene expression during development and environmental challenges in higher eukaryotes (Fujioka et al., 1999; Perry et al., 2010; Hnisz et al., 2015; Cannavò et al., 2016), especially given that, to my knowledge, no long-distance enhancer control elements have

been reported in the extremely gene-dense *D. discoideum* genome.

6.4 Buffering noise from nascent transcription in protein production

Having seen considerable differences in the levels of noise measured during nascent transcription, I then assessed how this changes during later steps of protein production. Compared to nascent transcription (Fig. 4.6), the 'noise differential', i.e. the difference in population variability as measured by the squared coefficient of variation (C_v^2), between actin genes is reduced in distributions of processed, cytoplasmic mRNA counts (Fig. 4.17), and is effectively lost when comparing variability of protein level distributions (Fig. 4.25). This suggests that variability in actin gene expression is primarily controlled post-transcriptionally, presumably to dampen noise arising from bursty transcription of *act5* and *act6* (as opposed to increasing noise in *act8* expression). Several studies have identified translation as the dominant factor in gene expression control (Thattai & van Oudenaarden, 2001; Schwanhäusser et al., 2011). Indeed, Fraser et al. (2004) showed that genes for which variable protein production might be harmful, such as essential genes or protein complex subunits, utilise a strategy of high transcription and low translation levels to reduce noise in the system. Measurement of translation and protein degradation would be useful for a more complete understanding of the mechanisms regulating the noise transmission in this system. Despite lacking a complete parameterisation of the process, this work directly addresses how transcriptional bursting behaviours occurring on the order of minutes affect steady-state mRNA and protein levels, the half-lives of which are typically on the order of hours. Live imaging and analysis of protein level fluctuations in individual cells could further inform these ideas.

If transcriptional noise is functional, as suggested above, then how can one reconcile this with the observation of reduced variability in later stages

of protein production? While these two conclusions may appear contradictory, I would argue that they are indeed compatible. The functional relevance of variable gene expression to environmental responses is at least twofold: (1) it enables 'bet-hedging' within a population by variable expression of a stress response factor facilitating survival of the portion of the population above or below a certain threshold; (2) it is associated with increased 'plasticity' in gene expression, that is, changes in average gene expression levels according to environmental conditions (Lehner, 2010). Identifying reduced noise in protein distributions compared to nascent transcription essentially rules out the former, as this assumes that variable transcription translates into variable protein levels within a population. While noise, in general, is strongly linked to plasticity of gene expression (Singh, 2013; Wolf et al., 2015) this relationship can be uncoupled for certain classes of genes (Lehner, 2010). Any mechanism behind such a relationship is unclear, and noise may even occur simply as a byproduct of gene expression plasticity (Lehner, 2010; Bajić & Poyatos, 2012). Given that plasticity concerns changes in average expression level, and not variability within a population, the fact that transcriptional noise is reduced at the protein level in this system would not affect the interpretation here. Therefore, variable actin gene expression could simply be a feature of plastic responses to environmental changes. This cannot be determined from the data presented here given that protein levels were only measured in a single environmental condition. A simple experiment measuring protein levels by flow cytometry at multiple culture densities for each actin gene, coupled with further experiments on nascent transcription in these conditions, could determine the likelihood of such a relationship in this system.

Bibliography

Addinall, S. G. & Lutkenhaus, J. (1996). FtsA is localized to the septum in an FtsZ-dependent manner. *Journal of bacteriology*, 178(24), 7167–7172.

Adler, M., Anjum, M., Berg, O. G., Andersson, D. I., & Sandegren, L. (2014). High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms. *Molecular biology and evolution*, 31(6), 1526–1535.

Alton, T. H. & Brenner, M. (1979). Comparison of proteins synthesized by anterior and posterior regions of Dictyostelium discoideum pseudoplasmodia. *Developmental biology*, 71(1), 1–7.

Alton, T. H. & Lodish, H. F. (1977). Developmental changes in messenger RNAs and protein synthesis in Dictyostelium discoideum. *Developmental biology*, 60(1), 180–206.

An, S. S., Möpps, B., Weber, K., & Bhattacharya, D. (1999). The origin and evolution of green algal and plant actins. *Molecular biology and evolution*, 16(2), 275–285.

Annesley, S. J. & Fisher, P. R. (2009). Dictyostelium discoideum–a model for many reasons. *Molecular and cellular biochemistry*, 329(1-2), 73–91.

Antolović, V., Miermont, A., Corrigan, A. M., & Chubb, J. R. (2017). Generation of Single-Cell Transcript Variability by Repression. *Current biology*, 27(12), 1811–1817.e3.

Artemenko, Y., Lampert, T. J., & Devreotes, P. N. (2014). Moving towards a paradigm: common mechanisms of chemotactic signaling in Dictyostelium and mammalian leukocytes. *Cellular and molecular life sciences*, 71(19), 3711–3747.

Bahar Halpern, K., Caspi, I., Lemze, D., Levy, M., Landen, S., Elinav, E., Ulitsky, I., & Itzkovitz, S. (2015a). Nuclear Retention of mRNA in Mammalian Tissues. *Cell reports*, 13(12), 2653–2662.

Bahar Halpern, K., Tanami, S., Landen, S., Chapal, M., Szlak, L., Hutzler, A., Nizhberg, A., & Itzkovitz, S. (2015b). Bursty gene expression in the intact mammalian liver. *Molecular cell*, 58(1), 147–156.

Bai, L., Charvin, G., Siggia, E. D., & Cross, F. R. (2010). Nucleosome-depleted regions in cell-cycle-regulated promoters ensure reliable gene expression in every cell cycle.

Developmental cell, 18(4), 544–555.

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., & Eichler, E. E. (2002). Recent Segmental Duplications in the Human Genome. *Science*, 297(5583), 1003–1007.

Baird, W. V. & Meagher, R. B. (1987). A complex gene superfamily encodes actin in petunia. *The EMBO journal*, 6(11), 3223–3231.

Bajić, D. & Poyatos, J. F. (2012). Balancing noise and plasticity in eukaryotic gene expression. *BMC genomics*, 13(1), 343.

Balázsi, G., van Oudenaarden, A., & Collins, J. J. (2011). Cellular decision making and biological noise: from microbes to mammals. *Cell*, 144(6), 910–925.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., & Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. *Nature genetics*, 38(6), 636–643.

Barklis, E. & Lodish, H. F. (1985). In situ localization of actin mRNA in Dictyostelium discoideum aggregates. *Experimental Cell Research*, 159(2), 479–486.

Basu, S., Fey, P., Pandit, Y., Dodson, R., Kibbe, W. A., & Chisholm, R. L. (2013). dictyBase 2013: integrating multiple Dictyostelid species. *Nucleic acids research*, 41(D1), D676–D683.

Batada, N. N., Urrutia, A. O., & Hurst, L. D. (2007). Chromatin remodelling is a major source of coexpression of linked genes in yeast. *Trends in genetics*, 23(10), 480–484.

Batenchuk, C., St-Pierre, S., Tepliakova, L., Adiga, S., Szuto, A., Kabbani, N., Bell, J. C., Baetz, K., & Kærn, M. (2011). Chromosomal position effects are linked to sir2-mediated variation in transcriptional burst size. *Biophysical journal*, 100(10), L56–8.

Battich, N., Stoeger, T., & Pelkmans, L. (2015). Control of Transcript Variability in Single Mammalian Cells. *Cell*, 163(7), 1596–1610.

Belyantseva, I. A., Perrin, B. J., Sonnemann, K. J., Zhu, M., Stepanyan, R., McGee, J., Frolenkov, G. I., Walsh, E. J., Friderici, K. H., Friedman, T. B., & Ervasti, J. M. (2009). Gamma-actin is required for cytoskeletal maintenance but not development. *Proceedings of the National Academy of Sciences*, 106(24), 9703–9708.

Ben-Shem, A., Garreau de Loubresse, N., Melnikov, S., Jenner, L., Yusupova, G., & Yusupov, M. (2011). The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science*, 334(6062), 1524–1529.

Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S. M., Singer, R. H., & Long, R. M. (1998). Localization of ASH1 mRNA particles in living yeast. *Molecular cell*, 2(4), 437–445.

Birchler, J. A. & Veitia, R. A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sci-*

ences, 109(37), 14746–14753.

Blake, W. J., Balázsi, G., Kohanski, M. A., Isaacs, F. J., Murphy, K. F., Kuang, Y., Cantor, C. R., Walt, D. R., & Collins, J. J. (2006). Phenotypic consequences of promoter-mediated transcriptional noise. *Molecular cell*, 24(6), 853–865.

Blanchoin, L., Boujemaa-Paterski, R., Sykes, C., & Plastino, J. (2014). Actin dynamics, architecture, and mechanics in cell motility. *Physiological reviews*, 94(1), 235–263.

Blaskovics, J. C. & Raper, K. B. (1957). Encystment stages of *Dictyostelium*. *The Biological Bulletin*, 113(1), 58–88.

Bloomfield, G. (2013). Sex in Dictyostelia. In *Dictyostelids* (pp. 129–148). Berlin, Heidelberg: Springer Berlin Heidelberg.

Bloomfield, G., Skelton, J., Ivens, A., Tanaka, Y., & Kay, R. R. (2010). Sex determination in the social amoeba Dictyostelium discoideum. *Science*, 330(6010), 1533–1536.

Bloomfield, G., Traynor, D., Sander, S. P., Veltman, D. M., Pachebat, J. A., & Kay, R. R. (2015). Neurofibromin controls macropinocytosis and phagocytosis in Dictyostelium. *eLife*, 4.

Boer, D. R., Freire-Rios, A., van den Berg, W. A. M., Saaki, T., Manfield, I. W., Kepinski, S., López-Vidrieo, I., Franco-Zorrilla, J. M., de Vries, S. C., Solano, R., Weijers, D., & Coll, M. (2014). Structural basis for DNA binding specificity by the auxin-dependent ARF transcription factors. *Cell*, 156(3), 577–589.

Bonner, J. T. (1947). Evidence for the formation of cell aggregates by chemotaxis in the development of the slime mold Dictyostelium discoideum. *The Journal of experimental zoology*, 106(1), 1–26.

Bonner, J. T. & Frascella, E. B. (1953). Variations in cell size during the development of the slime mold, Dictyostelium discoideum. *The Biological Bulletin*, 104(3), 297–300.

Bonner, J. T. & Slifkin, M. K. (1949). A Study of the Control of Differentiation: The Proportions of Stalk and Spore Cells in the Slime Mold Dictyostelium discoideum. *American Journal of Botany*, 36(10), 727.

Botstein, D. & Fink, G. R. (2011). Yeast: an experimental organism for 21st Century biology. *Genetics*, 189(3), 695–704.

Bowers, S. R., Calero-Nieto, F. J., Valeaux, S., Fernandez-Fuentes, N., & Cockerill, P. N. (2010). Runx1 binds as a dimeric complex to overlapping Runx1 sites within a palindromic element in the human GM-CSF enhancer. *Nucleic acids research*, 38(18), 6124–6134.

Bray, D. & Thomas, C. (1975). The actin content of fibroblasts. *Biochemical Journal*, 147(2), 221–228.

Bridges, C. B. (1936). The Bar "Gene" A Duplication. *Science*, 83(2148), 210–211.

Brivanlou, A. H. & Darnell, J. E. (2002). Signal transduction and the control of gene expres-

sion. *Science*, 295(5556), 813–818.

Brown, A. J. L. & Ish-Horowicz, D. (1981). Evolution of the 87A and 87C heat-shock loci in Drosophila. *Nature*, 290(5808), 677–682.

Brown, C. J., Todd, K. M., & Rosenzweig, R. F. (1998). Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Molecular biology and evolution*, 15(8), 931–942.

Brown, C. R., Mao, C., Falkovskaia, E., Jurica, M. S., & Boeger, H. (2013). Linking stochastic fluctuations in chromatin structure and gene expression. *PLoS biology*, 11(8), e1001621.

Brown, D. D., Wensink, P. C., & Jordan, E. (1972). A comparison of the ribosomal DNA's of Xenopus laevis and Xenopus mulleri: the evolution of tandem genes. *Journal of molecular biology*, 63(1), 57–73.

Brown, J. M. & Firtel, R. A. (1999). Regulation of cell-fate determination in Dictyostelium. *Developmental biology*, 216(2), 426–441.

Brown, K. M. & Gilmartin, G. M. (2003). A mechanism for the regulation of pre-mRNA 3' processing by human cleavage factor Im. *Molecular cell*, 12(6), 1467–1476.

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M. V., & Komar, A. A. (2016). Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Molecular cell*, 61(3), 341–351.

Bunnell, T. M., Burbach, B. J., Shimizu, Y., & Ervasti, J. M. (2011). β-Actin specifically controls cell growth, migration, and the G-actin pool. *Molecular biology of the cell*, 22(21), 4047–4058.

Bunnell, T. M. & Ervasti, J. M. (2010). Delayed embryonic development and impaired cell growth and survival in Actg1 null mice. *Cytoskeleton (Hoboken, N.J.)*, 67(9), 564–572.

Cai, J. J., Smith, D. K., Xia, X., & Yuen, K.-y. (2005). MBEToolbox: a Matlab toolbox for sequence data analysis in molecular biology and evolution. *BMC bioinformatics*, 6(1), 64.

Cairns, B. R. (2009). The logic of chromatin architecture and remodelling at promoters. *Nature*, 461(7261), 193–198.

Cannavò, E., Khoueiry, P., Garfield, D. A., Geeleher, P., Zichner, T., Gustafson, E. H., Ciglar, L., Korbel, J. O., & Furlong, E. E. M. (2016). Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Current biology*, 26(1), 38–51.

Carey, L. B., van Dijk, D., Sloot, P. M. A., Kaandorp, J. A., & Segal, E. (2013). Promoter sequence determines the relationship between expression level and noise. *PLoS biology*, 11(4), e1001528.

Chaney, J. L. & Clark, P. L. (2015). Roles for Synonymous Codon Usage in Protein Biogenesis. *Annual review of biophysics*, 44(1), 143–166.

Chang, D. & Duda, T. F. (2012). Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. *Molecular biology and evolution*, 29(8), 2019–2029.

Chang, G. S., Noegel, A. A., Mavrich, T. N., Müller, R., Tomsho, L., Ward, E., Felder, M., Jiang, C., Eichinger, L., Glöckner, G., Schuster, S. C., & Pugh, B. F. (2012). Unusual combinatorial involvement of poly-A/T tracts in organizing genes and chromatin in Dictyostelium. *Genome research*, 22(6), 1098–1106.

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., & Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. *Nature reviews genetics*, 8(10), 762–775.

Chisholm, R. L. & Firtel, R. A. (2004). Insights into morphogenesis from a simple developmental system. *Nature reviews molecular cell biology*, 5(7), 531–541.

Chlebowski, A., Lubas, M., Jensen, T. H., & Dziembowski, A. (2013). RNA decay machines: the exosome. *Biochimica et biophysica acta*, 1829(6-7), 552–560.

Chowdhary, R., Ali, R. A., ALBIG, W., Doenecke, D., & Bajic, V. B. (2005). Promoter modeling: the case study of mammalian histone promoters. *Bioinformatics (Oxford, England)*, 21(11), 2623–2628.

Chubb, J. R., Stevense, M., Cannon, D., Muramoto, T., & Corrigan, A. M. (2013). Imaging nascent RNA dynamics in Dictyostelium. *Methods in molecular biology*, 1042, 101–113.

Chubb, J. R., Trcek, T., Shenoy, S. M., & Singer, R. H. (2006). Transcriptional pulsing of a developmental gene. *Current biology*, 16(10), 1018–1025.

Clarke, T. F. & Clark, P. L. (2010). Increased incidence of rare codon clusters at 5' and 3' gene termini: implications for function. *BMC genomics*, 11(1), 118.

Cohen, S. M., Knecht, D., Lodish, H. F., & Loomis, W. F. (1986). DNA sequences required for expression of a Dictyostelium actin gene. *The EMBO journal*, 5(12), 3361–3366.

Cohn, R. H. & Kedes, L. H. (1979). Nonallelic histone gene clusters of individual sea urchins (Lytechinus pictus): polarity and gene organization. *Cell*, 18(3), 843–853.

Coloma, A. & Lodish, H. F. (1981). Synthesis of spore- and stalk-specific proteins during differentiation of Dictyostelium discoideum. *Developmental biology*, 81(2), 238–244.

Condeelis, J. & Singer, R. H. (2005). How and why does beta-actin mRNA target? *Biology of the cell*, 97(1), 97–110.

Corrigan, A. M. & Chubb, J. R. (2014). Regulation of transcriptional bursting by a naturally oscillating signal. *Current biology*, 24(2), 205–211.

Corrigan, A. M., Tunnacliffe, E., Cannon, D., & Chubb, J. R. (2016). A continuum model of transcriptional bursting. *eLife*, 5, 631.

Crawford, K., Flick, R., Close, L., Shelly, D., Paul, R., Bove, K., Kumar, A., & Lessard, J. (2002). Mice lacking skeletal muscle actin show reduced muscle strength and growth

deficits and die during the neonatal period. *Molecular and cellular biology*, 22(16), 5887–5896.

Dar, R. D., Razooky, B. S., Singh, A., Trimeloni, T. V., McCollum, J. M., Cox, C. D., Simpson, M. L., & Weinberger, L. S. (2012). Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences*, 109(43), 17454–17459.

Datta, S. & Firtel, R. A. (1988). An 80-bp cis-acting regulatory region controls cAMP and development regulation of a prestalk gene in Dictyostelium. *Genes & development*, 2(3), 294–304.

Davoli, T. & de Lange, T. (2011). The causes and consequences of polyploidy in normal development and cancer. *Annual review of cell and developmental biology*, 27(1), 585–610.

De Lozanne, A. & Spudich, J. A. (1987). Disruption of the Dictyostelium myosin heavy chain gene by homologous recombination. *Science*, 236(4805), 1086–1091.

Dean, E. J., Davis, J. C., Davis, R. W., & Petrov, D. A. (2008). Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS genetics*, 4(7), e1000113.

DeLuna, A., Springer, M., Kirschner, M. W., & Kishony, R. (2010). Need-based up-regulation of protein levels in response to deletion of their duplicate genes. *PLoS biology*, 8(3), e1000347.

DeLuna, A., Vetsigian, K., Shoresh, N., Hegreness, M., Colón-González, M., Chao, S., & Kishony, R. (2008). Exposing the fitness contribution of duplicated genes. *Nature genetics*, 40(5), 676–681.

Derti, A., Garrett-Engele, P., Macisaac, K. D., Stevens, R. C., Sriram, S., Chen, R., Rohl, C. A., Johnson, J. M., & Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. *Genome research*, 22(6), 1173–1183.

Dey, S. S., Foley, J. E., Limsirichai, P., Schaffer, D. V., & Arkin, A. P. (2015). Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. *Molecular systems biology*, 11(5), 806–806.

Dobzhansky, T. (1932). The Baroid Mutation in Drosophila Melanogaster. *Genetics*, 17(4), 369–392.

Doi, M., Wachi, M., Ishino, F., Tomioka, S., Ito, M., Sakagami, Y., Suzuki, A., & Matsuhashi, M. (1988). Determinations of the DNA sequence of the mreB gene and of the gene products of the mre region that function in formation of the rod shape of Escherichia coli cells. *Journal of bacteriology*, 170(10), 4619–4624.

Dong, D., Yuan, Z., & Zhang, Z. (2011). Evidences for increased expression variation of duplicate genes in budding yeast: from cis- to trans-regulation effects. *Nucleic acids research*, 39(3), 837–847.

Dugina, V., Zwaenepoel, I., Gabbiani, G., Clément, S., & Chaponnier, C. (2009). β- and γ-cytoplasmic actins display distinct distribution and functional diversity. *Journal of cell science*, 122(16), 2980–2988.

Eichinger, L., Pachebat, J. A., Glöckner, G., Rajandream, M.-A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., Tunggal, B., Kummerfeld, S., Madera, M., Konfortov, B. A., Rivero, F., Bankier, A. T., Lehmann, R., Hamlin, N., Davies, R., Gaudet, P., Fey, P., Pilcher, K., Chen, G., Saunders, D., Sodergren, E., Davis, P., Kerhornou, A., Nie, X., Hall, N., Anjard, C., Hemphill, L., Bason, N., Farbrother, P., Desany, B., Just, E., Morio, T., Rost, R., Churcher, C., Cooper, J., Haydock, S., van Driessche, N., Cronin, A., Goodhead, I., Muzny, D., Mourier, T., Pain, A., Lu, M., Harper, D., Lindsay, R., Hauser, H., James, K., Quiles, M., Madan Babu, M., Saito, T., Buchrieser, C., Wardroper, A., Felder, M., Thangavelu, M., Johnson, D., Knights, A., Loulseged, H., Mungall, K., Oliver, K., Price, C., Quail, M. A., Urushihara, H., Hernandez, J., Rabbinowitsch, E., Steffen, D., Sanders, M., Ma, J., Kohara, Y., Sharp, S., Simmonds, M., Spiegler, S., Tivey, A., Sugano, S., White, B., Walker, D., Woodward, J., Winckler, T., Tanaka, Y., Shaulsky, G., Schleicher, M., Weinstock, G., Rosenthal, A., Cox, E. C., Chisholm, R. L., Gibbs, R., Loomis, W. F., Platzer, M., Kay, R. R., Williams, J., Dear, P. H., Noegel, A. A., Barrell, B., & Kuspa, A. (2005). The genome of the social amoeba Dictyostelium discoideum. *Nature*, 435(7038), 43–57.

Eichinger, L. & Rivero, F. (2006). *Dictyostelium discoideum protocols*, volume 346. Springer.

Eichinger, L. & Rivero, F. (2013). Dictyostelium discoideum Protocols. Humana Press, Totowa, NJ.

Eickbush, T. H. & Eickbush, D. G. (2007). Finely Orchestrated Movements: Evolution of the Ribosomal RNA Genes. *Genetics*, 175(2), 477–485.

Eirín-López, J. M., González-Tizón, A. M., Martínez, A., & Méndez, J. (2004). Birth-and-death evolution with strong purifying selection in the histone H1 multigene family and the origin of orphon H1 genes. *Molecular biology and evolution*, 21(10), 1992–2003.

Eirín-López, J. M., Rebordinos, L., Rooney, A. P., & Rozas, J. (2012). The birth-and-death evolution of multigene families revisited. *Genome dynamics*, 7, 170–196.

Elkon, R., Ugalde, A. P., & Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. *Nature reviews genetics*, 14(7), 496–506.

Elowitz, M. B., Levine, A. J., Siggia, E. D., & Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584), 1183–1186.

Falvo, J. V., Lin, C. H., Tsytsykova, A. V., Hwang, P. K., Thanos, D., Goldfeld, A. E., & Maniatis, T. (2008). A dimer-specific function of the transcription factor NFATp. *Proceedings of the National Academy of Sciences*, 105(50), 19637–19642.

Femino, A. M., Fay, F. S., Fogarty, K., & Singer, R. H. (1998). Visualization of single RNA transcripts in situ. *Science*, 280(5363), 585–590.

Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I. K., Sharon, E., Lubling, Y., Widom, J., & Segal, E. (2008). Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS computational biology*, 4(11), e1000216.

Filosa, M. F. & Dengler, R. E. (1972). Ultrastructure of macrocyst formation in the cellular slime mold, Dictyostelium mucoroides: Extensive phagocytosis of amoebae by a specialized cell. *Developmental biology*, 29(1), 1–16.

Firtel, R. A., Baxter, L., & Lodish, H. F. (1973). Actinomycin D and the regulation of enzyme biosynthesis during development of Dictyostelium discoideum. *Journal of molecular biology*, 79(2), 315–327.

Firtel, R. A., Timm, R., Kimmel, A. R., & McKeown, M. (1979). Unusual nucleotide sequences at the 5' end of actin genes in Dictyostelium discoideum. *Proceedings of the National Academy of Sciences of the United States of America*, 76(12), 6206–6210.

Francino, M. P. (2005). An adaptive radiation model for the origin of new gene functions. *Nature genetics*, 37(6), 573–577.

Fraser, H. B., Hirsh, A. E., Giaever, G., Kumm, J., & Eisen, M. B. (2004). Noise minimization in eukaryotic gene expression. *PLoS biology*, 2(6), e137.

Fraser, P. & Bickmore, W. (2007). Nuclear organization of the genome and the potential for gene regulation. *Nature*, 447(7143), 413–417.

Friedl, P., Borgmann, S., & Bröcker, E. B. (2001). Amoeboid leukocyte crawling through extracellular matrix: lessons from the Dictyostelium paradigm of cell movement. *Journal of leukocyte biology*, 70(4), 491–509.

Fu, J., Murphy, K. A., Zhou, M., Li, Y. H., Lam, V. H., Tabuloc, C. A., Chiu, J. C., & Liu, Y. (2016). Codon usage affects the structure and function of the Drosophila circadian clock protein PERIOD. *Genes & development*, 30(15), 1761–1775.

Fujioka, M., Emi-Sarker, Y., Yusibova, G. L., Goto, T., & Jaynes, J. B. (1999). Analysis of an even-skipped rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients. *Development (Cambridge, England)*, 126(11), 2527–2538.

Futuyma, D. J. (2005). *Evolution*. Evolution. Sinauer Associates.

Ghosh, G., van Duyne, G., Ghosh, S., & Sigler, P. B. (1995). Structure of NF-kappa B p50 homodimer bound to a kappa B site. *Nature*, 373(6512), 303–310.

Golding, I., Paulsson, J., Zawilski, S. M., & Cox, E. C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6), 1025–1036.

Golé, L., Rivière, C., Hayakawa, Y., & Rieu, J.-P. (2011). A quorum-sensing factor in vegetative Dictyostelium discoideum cells revealed by quantitative migration analysis. *PloS*

one, 6(11), e26901.

Gowing, L. R., Tellam, R. L., & Banyard, M. R. (1984). Microfilament organization and total actin content are decreased in hybrids derived from the fusion of HeLa cells with human fibroblasts. *Journal of cell science*, 69, 137–146.

Green, M. M. (1963). Unequal crossing over and the genetical organization of the white locus of Drosophila melanogaster. *Zeitschrift für Vererbungslehre*, 94(2), 200–214.

Greer, S., Honeywell, R., Geletu, M., Arulanandam, R., & Raptis, L. (2010). Housekeeping genes; expression levels may change with density of cultured cells. *Journal of immunological methods*, 355(1-2), 76–79.

Gunning, P. W., Ghoshdastider, U., Whitaker, S., Popp, D., & Robinson, R. C. (2015). The evolution of compositionally and functionally distinct actin filaments. *Journal of cell science*, 128(11), 2009–2019.

Hakes, L., Pinney, J. W., Lovell, S. C., Oliver, S. G., & Robertson, D. L. (2007). All duplicates are not equal: the difference between small-scale and genome duplication. *Genome biology*, 8(10), R209.

Hanes, S. D. & Brent, R. (1989). DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9. *Cell*, 57(7), 1275–1283.

Hanson, J. & Huxley, H. E. (1957). Quantitative studies on the structure of cross-striated myofibrils. II. Investigations by biochemical techniques. *Biochimica et biophysica acta*, 23(2), 250–260.

Hardison, R. C. (2012). Evolution of Hemoglobin and Its Genes. *Cold Spring Harbor perspectives in medicine*, 2(12), a011627–a011627.

Hershberg, R., Yeger-Lotem, E., & Margalit, H. (2005). Chromosomal organization is shaped by the transcription regulatory network. *Trends in genetics*, 21(3), 138–142.

Hjorth, A., Datta, S., Khanna, N. C., & Firtel, R. A. (1988). Analysis of cis and trans elements involved in cAMP-inducible gene expression in Dictyostelium discoideum. *Developmental genetics*, 9(4-5), 435–454.

Hjorth, A. L., Khanna, N. C., & Firtel, R. A. (1989). A trans-acting factor required for cAMP-induced gene expression in Dictyostelium is regulated developmentally and induced by cAMP. *Genes & development*, 3(6), 747–759.

Hjorth, A. L., Pears, C., Williams, J. G., & Firtel, R. A. (1990). A developmentally regulated trans-acting factor recognizes dissimilar G/C-rich elements controlling a class of cAMP-inducible Dictyostelium genes. *Genes & development*, 4(3), 419–432.

Hnisz, D., Schuijers, J., Lin, C. Y., Weintraub, A. S., Abraham, B. J., Lee, T. I., Bradner, J. E., & Young, R. A. (2015). Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Molecular cell*, 58(2), 362–370.

Hocine, S., Vera, M., Zenklusen, D., & Singer, R. H. (2015). Promoter-Autonomous Func-

tioning in a Controlled Environment using Single Molecule FISH. *Scientific reports*, 5(1), 9934.

Holmes, W. F., Braastad, C. D., Mitra, P., Hampe, C., Doenecke, D., ALBIG, W., Stein, J. L., van Wijnen, A. J., & Stein, G. S. (2005). Coordinate control and selective expression of the full complement of replication-dependent histone H4 genes in normal and cancer cells. *The Journal of biological chemistry*, 280(45), 37400–37407.

Holt, C. A. & Childs, G. (1984). A new family of tandem repetitive early histone genes in the sea urchin Lytechinus pictus: evidence for concerted evolution within tandem arrays. *Nucleic acids research*, 12(16), 6455–6471.

Hori, R. & Firtel, R. A. (1994). Identification and characterization of multiple A/T-rich cis-acting elements that control expression from Dictyostelium actin promoters: the Dictyostelium actin upstream activating sequence confers growth phase expression and has enhancer-like properties. *Nucleic acids research*, 22(23), 5099–5111.

Hornung, G., Bar-Ziv, R., Rosin, D., Tokuriki, N., Tawfik, D. S., Oren, M., & Barkai, N. (2012). Noise-mean relationship in mutated promoters. *Genome research*, 22(12), 2409–2417.

Hu, J., Lutz, C. S., Wilusz, J., & Tian, B. (2005). Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA*, 11(10), 1485–1493.

Huang, S., McDowell, J. M., Weise, M. J., & Meagher, R. B. (1996). The Arabidopsis profilin gene family. Evidence for an ancient split between constitutive and pollen-specific profilin genes. *Plant physiology*, 111(1), 115–126.

Hughes, A. L. & Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335(6186), 167–170.

Hughes, A. L. & Nei, M. (1989a). Evolution of the major histocompatibility complex: independent origin of nonclassical class I genes in different groups of mammals. *Molecular biology and evolution*, 6(6), 559–579.

Hughes, A. L. & Nei, M. (1989b). Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proceedings of the National Academy of Sciences of the United States of America*, 86(3), 958–962.

Hughes, T. & Liberles, D. A. (2008). The power-law distribution of gene family size is driven by the pseudogenisation rate's heterogeneity between gene families. *Gene*, 414(1-2), 85–94.

Hüttelmaier, S., Zenklusen, D., Lederer, M., Dictenberg, J., Lorenz, M., Meng, X., Bassell, G. J., Condeelis, J., & Singer, R. H. (2005). Spatial regulation of beta-actin translation by Src-dependent phosphorylation of ZBP1. *Nature*, 438(7067), 512–515.

Huynen, M. A. & van Nimwegen, E. (1998). The frequency distribution of gene family sizes in complete genomes. *Molecular biology and evolution*, 15(5), 583–589.

Ingram, V. M. (1961). Gene evolution and the haemoglobins. *Nature*, 189, 704–708.

Innan, H. & Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature reviews genetics*, 11(2), 97–108.

Iranfar, N., Fuller, D., & Loomis, W. F. (2006). Transcriptional regulation of post-aggregation genes in Dictyostelium by a feed-forward loop involving GBF and LagC. *Developmental biology*, 290(2), 460–469.

Jaeger, M. A., Sonnemann, K. J., Fitzsimons, D. P., Prins, K. W., & Ervasti, J. M. (2009). Context-dependent functional substitution of alpha-skeletal actin by gamma-cytoplasmic actin. *FASEB journal*, 23(7), 2205–2214.

Jansen, R., Bussemaker, H. J., & Gerstein, M. (2003). Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. *Nucleic acids research*, 31(8), 2242–2251.

Jensen, R. B. & Gerdes, K. (1997). Partitioning of plasmid R1. The ParM protein exhibits ATPase activity and interacts with the centromere-like ParR-parC complex. *Journal of molecular biology*, 269(4), 505–513.

Joseph, J. M., Fey, P., Ramalingam, N., Liu, X. I., Rohlfs, M., Noegel, A. A., Müller-Taubenberger, A., Glöckner, G., & Schleicher, M. (2008). The actinome of Dictyostelium discoideum in comparison to actins and actin-related proteins from other organisms. *PloS one*, 3(7), e2654.

Kafri, R., Bar-Even, A., & Pilpel, Y. (2005). Transcription control reprogramming in genetic backup circuits. *Nature genetics*, 37(3), 295–299.

Kafri, R., Levy, M., & Pilpel, Y. (2006). The regulatory utilization of genetic redundancy through responsive backup circuits. *Proceedings of the National Academy of Sciences of the United States of America*, 103(31), 11653–11658.

Kandasamy, M. K., Burgos-Rivera, B., McKinney, E. C., Ruzicka, D. R., & Meagher, R. B. (2007). Class-specific interaction of profilin and ADF isovariants with actin in the regulation of plant development. *The Plant cell*, 19(10), 3111–3126.

Kandasamy, M. K., McKinney, E. C., & Meagher, R. B. (2009). A single vegetative actin isovariant overexpressed under the control of multiple regulatory sequences is sufficient for normal Arabidopsis development. *The Plant cell*, 21(3), 701–718.

Kane, J. F. (1995). Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. *Current opinion in biotechnology*, 6(5), 494–500.

Katju, V. & Lynch, M. (2006). On the formation of novel genes by duplication in the Caenorhabditis elegans genome. *Molecular biology and evolution*, 23(5), 1056–1067.

Kay, R. R., Langridge, P., Traynor, D., & Hoeller, O. (2008). Changing directions in the study of chemotaxis. *Nature reviews molecular cell biology*, 9(6), 455–463.

Kay, R. R. & Thompson, C. R. L. (2009). Forming Patterns in Development without Morphogen Gradients: Scattered Differentiation and Sorting Out. *Cold Spring Harbor perspectives in biology*, 1(6), a001503–a001503.

Kazazian, H. H. (2004). Mobile elements: drivers of genome evolution. *Science*, 303(5664), 1626–1632.

Kempe, H., Schwabe, A., Crémazy, F., Verschure, P. J., & Bruggeman, F. J. (2015). The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise. *Molecular biology of the cell*, 26(4), 797–804.

Kenmochi, N., Kawaguchi, T., Rozen, S., Davis, E., Goodman, N., Hudson, T. J., Tanaka, T., & Page, D. C. (1998). A map of 75 human ribosomal protein genes. *Genome research*, 8(5), 509–523.

Khorasanizadeh, S. & Rastinejad, F. (2001). Nuclear-receptor interactions on DNA-response elements. *Trends in biochemical sciences*, 26(6), 384–390.

Kim, S. J., Yoon, J. S., Shishido, H., Yang, Z., Rooney, L. A., Barral, J. M., & Skach, W. R. (2015). Translational tuning optimizes nascent protein folding in cells. *Science*, 348(6233), 444–448.

Kimmel, A. R. & Firtel, R. A. (1983). Sequence organization in Dictyostelium: unique structure at the 5'-ends of protein coding genes. *Nucleic acids research*, 11(2), 541–552.

Kindle, K. L. & Firtel, R. A. (1978). Identification and analysis of Dictyostelium actin genes, a family of moderately repeated genes. *Cell*, 15(3), 763–778.

Kislauskis, E. H. (1994). Sequences responsible for intracellular localization of beta-actin messenger RNA also affect cell phenotype. *The Journal of cell biology*, 127(2), 441–451.

Knecht, D. A., Cohen, S. M., Loomis, W. F., & Lodish, H. F. (1986). Developmental regulation of Dictyostelium discoideum actin gene fusions carried on low-copy and high-copy transformation vectors. *Molecular and cellular biology*, 6(11), 3973–3983.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. *Molecular cell*, 58(4), 610–620.

Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings. Biological sciences*, 279(1749), 5048–5057.

Kornberg, R. D. & Stryer, L. (1988). Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic acids research*, 16(14A), 6677–6690.

Koszul, R., Dujon, B., & Fischer, G. (2006). Stability of large segmental duplications in the yeast genome. *Genetics*, 172(4), 2211–2222.

Kovar, D. R., Sirotkin, V., & Lord, M. (2011). Three's company: the fission yeast actin cytoskeleton. *Trends in cell biology*, 21(3), 177–187.

Kramer, E. B. & Farabaugh, P. J. (2007). The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA*, 13(1), 87–96.

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. *Science*, 324(5924), 255–258.

Kumar, A., Crawford, K., Close, L., Madison, M., Lorenz, J., Doetschman, T., Pawlowski, S., Duffy, J., Neumann, J., Robbins, J., Boivin, G. P., O'Toole, B. A., & Lessard, J. L. (1997). Rescue of cardiac alpha-actin-deficient mice by enteric smooth muscle gamma-actin. *Proceedings of the National Academy of Sciences of the United States of America*, 94(9), 4406–4411.

Kunkel, G. R. & Martinson, H. G. (1981). Nucleosomes will not form on double-stranded RNa or over poly(dA).poly(dT) tracts in recombinant DNA. *Nucleic acids research*, 9(24), 6869–6888.

Kuspa, A. & Loomis, W. F. (1992). Tagging developmental genes in Dictyostelium by restriction enzyme-mediated integration of plasmid DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 89(18), 8803–8807.

Lahr, D. J. G., Nguyen, T. B., Barbero, E., & Katz, L. A. (2011). Evolution of the actin gene family in testate lobose amoebae (Arcellinida) is characterized by two distinct clades of paralogs and recent independent expansions. *Molecular biology and evolution*, 28(1), 223–236.

Lan, X. & Pritchard, J. K. (2016). Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science*, 352(6288), 1009–1013.

Lawrence, J. B., Singer, R. H., & Marselle, L. M. (1989). Highly localized tracks of specific transcripts within interphase nuclei visualized by in situ hybridization. *Cell*, 57(3), 493–502.

Lehner, B. (2010). Conflict between noise and plasticity in yeast. *PLoS genetics*, 6(11), e1001185.

Lenstra, T. L., Rodriguez, J., Chen, H., & Larson, D. R. (2016). Transcription Dynamics in Living Cells. *Annual review of biophysics*, 45(1), 25–47.

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2010a). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics (Oxford, England)*, 26(4), 493–500.

Li, J., Yuan, Z., & Zhang, Z. (2010b). The cellular robustness by genetic redundancy in budding yeast. *PLoS genetics*, 6(11), e1001187.

Lindås, A.-C., Valegård, K., & Ettema, T. J. G. (2017). Archaeal Actin-Family Filament Systems. In J. Löwe & L. A. Amos (Eds.), *Prokaryotic Cytoskeletons* (pp. 379–392). Cham: Springer International Publishing.

López-Maury, L., Marguerat, S., & Bähler, J. (2008). Tuning gene expression to changing

environments: from rapid responses to evolutionary adaptation. *Nature reviews genetics*, 9(8), 583–593.

Losick, R. & Desplan, C. (2008). Stochasticity and cell fate. *Science*, 320(5872), 65–68.

Luisi, B. F., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R., & Sigler, P. B. (1991). Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature*, 352(6335), 497–505.

Mandel, C. R., Bai, Y., & Tong, L. (2008). Protein factors in pre-mRNA 3´-end processing. *Cellular and molecular life sciences*, 65(7-8), 1099–1122.

Manstein, D. J., Titus, M. A., De Lozanne, A., & Spudich, J. A. (1989). Gene replacement in Dictyostelium: generation of myosin null mutants. *The EMBO journal*, 8(3), 923–932.

Margolskee, J. P. & Lodish, H. F. (1980). The regulation of the synthesis of actin and two other proteins induced early in Dictyostelium discoideum development. *Developmental biology*, 74(1), 50–64.

Marguerat, S. & Bähler, J. (2012). Coordinating genome expression with cell size. *Trends in genetics*, 28(11), 560–565.

Martinsohn, J. T., Sousa, A. B., Guethlein, L. A., & Howard, J. C. (1999). The gene conversion hypothesis of MHC evolution: a review. *Immunogenetics*, 50(3-4), 168–200.

Marzluff, W. F., Gongidi, P., Woods, K. R., Jin, J., & Maltais, L. J. (2002). The Human and Mouse Replication-Dependent Histone Genes. *Genomics*, 80(5), 487–498.

Matsuo, Y. & Yamazaki, T. (1989). Nucleotide variation and divergence in the histone multigene family in Drosophila melanogaster. *Genetics*, 122(1), 87–97.

McDowell, J. M., Huang, S., McKinney, E. C., An, Y. Q., & Meagher, R. B. (1996). Structure and evolution of the actin gene family in Arabidopsis thaliana. *Genetics*, 142(2), 587–602.

McKeown, M. & Firtel, R. A. (1981a). Differential expression and 5' end mapping of actin genes in Dictyostelium. *Cell*, 24(3), 799–807.

McKeown, M. & Firtel, R. A. (1981b). Evidence for sub-families of actin genes in Dictyostelium as determined by comparisons of 3' end sequences. *Journal of molecular biology*, 151(4), 593–606.

McKeown, M., Taylor, W. C., Kindle, K. L., Firtel, R. A., Bender, W., & Davidson, N. (1978). Multiple, heterogeneous actin genes in Dictyostelium. *Cell*, 15(3), 789–800.

McKinnon, P. J. & Burgoyne, L. A. (1984). Regulation of the cellular actin levels in response to changes in the cell density in ataxia telangiectasia lymphoblastoid cells. *Biochemical and biophysical research communications*, 119(2), 561–566.

Mehdy, M. C., Ratner, D., & Firtel, R. A. (1983). Induction and modulation of cell-type-specific gene expression in Dictyostelium. *Cell*, 32(3), 763–771.

Mendonça, A. G., Alves, R. J., & Pereira-Leal, J. B. (2011). Loss of genetic redundancy in

reductive genome evolution. *PLoS computational biology*, 7(2), e1001082.

Michalak, P. (2008). Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, 91(3), 243–248.

Miller, O. L. & Beatty, B. R. (1969). Visualization of nucleolar genes. *Science*, 164(3882), 955–957.

Millevoi, S. & Vagner, S. (2010). Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic acids research*, 38(9), 2757–2774.

Mohammed, H., Hernando-Herraez, I., Savino, A., Scialdone, A., Macaulay, I., Mulas, C., Chandra, T., Voet, T., Dean, W., Nichols, J., Marioni, J. C., & Reik, W. (2017). Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation. *Cell reports*, 20(5), 1215–1228.

Mueller, F., Senecal, A., Tantale, K., Marie-Nelly, H., Ly, N., Collin, O., Basyuk, E., Bertrand, E., Darzacq, X., & Zimmer, C. (2013). FISH-quant: automatic counting of transcripts in 3D FISH images. *Nature methods*, 10(4), 277–278.

Müller, C. W., Rey, F. A., Sodeoka, M., Verdine, G. L., & Harrison, S. C. (1995). Structure of the NF-kappa B p50 homodimer bound to DNA. *Nature*, 373(6512), 311–317.

Muramoto, T., Cannon, D., Gierliński, M., Corrigan, A., Barton, G. J., & Chubb, J. R. (2012). Live imaging of nascent RNA dynamics reveals distinct types of transcriptional pulse regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 109(19), 7350–7355.

Muramoto, T., Müller, I., Thomas, G., Melvin, A., & Chubb, J. R. (2010). Methylation of H3K4 Is required for inheritance of active transcriptional states. *Current biology*, 20(5), 397–406.

Musso, G., Costanzo, M., Huangfu, M., Smith, A. M., Paw, J., San Luis, B.-J., Boone, C., Giaever, G., Nislow, C., Emili, A., & Zhang, Z. (2008). The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome research*, 18(7), 1092–1099.

Nei, M. (1969). Gene Duplication and Nucleotide Substitution in Evolution. *Nature*, 221(5175), 40–42.

Nei, M., Gu, X., & Sitnikova, T. (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences of the United States of America*, 94(15), 7799–7806.

Nei, M., Rogozin, I. B., & Piontkivska, H. (2000). Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20), 10866–10871.

Nei, M. & Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annual review of genetics*, 39(1), 121–152.

Nellen, W., Silan, C., & Firtel, R. A. (1984). DNA-mediated transformation in Dictyostelium discoideum: regulated expression of an actin gene fusion. *Molecular and cellular biology*, 4(12), 2890–2898.

Nellen, W., Silan, C., Saur, U., & Firtel, R. A. (1986). Regulatory sequences in the promoter of the Dictyostelium Actin 6 gene. *The EMBO journal*, 5(12), 3367–3372.

Niimura, Y. & Nei, M. (2003). Evolution of olfactory receptor genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 100(21), 12235–12240.

Niimura, Y. & Nei, M. (2006). Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates. *Journal of human genetics*, 51(6), 505–517.

Noegel, A., Gerisch, G., Stadler, J., & Westphal, M. (1986). Complete sequence and transcript regulation of a cell adhesion protein from aggregating Dictyostelium cells. *The EMBO journal*, 5(7), 1473–1476.

Noegel, A. A. & Schleicher, M. (2000). The actin cytoskeleton of Dictyostelium: a story told by mutants. *Journal of cell science*, 113 (Pt 5), 759–766.

Nowak, K. J., Ravenscroft, G., Jackaman, C., Filipovska, A., Davies, S. M., Lim, E. M., Squire, S. E., Potter, A. C., Baker, E., Clément, S., Sewry, C. A., Fabian, V., Crawford, K., Lessard, J. L., Griffiths, L. M., Papadimitriou, J. M., Shen, Y., Morahan, G., Bakker, A. J., Davies, K. E., & Laing, N. G. (2009). Rescue of skeletal muscle alpha-actin-null mice by cardiac (fetal) alpha-actin. *The Journal of cell biology*, 185(5), 903–915.

Nowak, M. A., Boerlijst, M. C., Cooke, J., & Smith, J. M. (1997). Evolution of genetic redundancy. *Nature*, 388(6638), 167–171.

O'Day, D. H. (1979). Aggregation during sexual development in Dictyostelium discoideum. *Canadian journal of microbiology*, 25(12), 1416–1426.

Ohno, S. (1970). *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ohta, T. (1976). Simple model for treating evolution of multigene families. *Nature*, 263(5572), 74–76.

Ohta, T. (1983). On the evolution of multigene families. *Theoretical population biology*, 23(2), 216–240.

Padovan-Merhar, O., Nair, G. P., Biaesch, A. G., Mayer, A., Scarfone, S., Foley, S. W., Wu, A. R., Churchman, L. S., Singh, A., & Raj, A. (2015). Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Molecular cell*, 58(2), 339–352.

Pan, P., Hall, E. M., & Bonner, J. T. (1972). Folic acid as second chemotactic substance in the cellular slime moulds. *Nature*, 237(75), 181–182.

Pánek, T., Zadrobílková, E., Walker, G., Brown, M. W., Gentekaki, E., Hroudová, M., Kang,

S., Roger, A. J., Tice, A. K., Vlček, Č., & Čepička, I. (2016). First multigene analysis of Archamoebae (Amoebozoa: Conosa) robustly reveals its phylogeny and shows that Entamoebidae represents a deep lineage of the group. *Molecular Phylogenetics and Evolution*, 98, 41–51.

Papp, B., Pál, C., & Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424(6945), 194–197.

Parikh, A., Miranda, E. R., Katoh-Kurasawa, M., Fuller, D., Rot, G., Zagar, L., Curk, T., Sucgang, R., Chen, R., Zupan, B., Loomis, W. F., Kuspa, A., & Shaulsky, G. (2010). Conserved developmental transcriptomes in evolutionarily divergent species. *Genome biology*, 11(3), R35.

Pavlovic, J., Haribabu, B., & Dottin, R. P. (1989). Identification of a signal transduction response sequence element necessary for induction of a Dictyostelium discoideum gene by extracellular cyclic AMP. *Molecular and cellular biology*, 9(11), 4660–4669.

Pears, C. J. & Williams, J. G. (1988). Multiple copies of a G-rich element upstream of a cAMP-inducible Dictyostelium gene are necessary but not sufficient for efficient gene expression. *Nucleic acids research*, 16(17), 8467–8486.

Pechmann, S. & Frydman, J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature structural & molecular biology*, 20(2), 237–243.

Pelham, R. J. & Chang, F. (2002). Actin dynamics in the contractile ring during cytokinesis in fission yeast. *Nature*, 419(6902), 82–86.

Perrin, B. J. & Ervasti, J. M. (2010). The actin gene family: function follows isoform. *Cytoskeleton (Hoboken, N.J.)*, 67(10), 630–634.

Perry, M. W., Boettiger, A. N., Bothma, J. P., & Levine, M. (2010). Shadow enhancers foster robustness of Drosophila gastrulation. *Current biology*, 20(17), 1562–1567.

Petes, T. D. (1980). Unequal meiotic recombination within tandem arrays of yeast ribosomal DNA genes. *Cell*, 19(3), 765–774.

Piontkivska, H., Rooney, A. P., & Nei, M. (2002). Purifying Selection and Birth-and-death Evolution in the Histone H4 Gene Family. *Molecular biology and evolution*, 19(5), 689–697.

Pollard, T. D. & Borisy, G. G. (2003). Cellular motility driven by assembly and disassembly of actin filaments. *Cell*, 112(4), 453–465.

Prabhu, Y. & Eichinger, L. (2006). The Dictyostelium repertoire of seven transmembrane domain receptors. *European journal of cell biology*, 85(9-10), 937–946.

Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K. E., Graveley, B. R., & Coller, J. (2015). Codon optimality is a major determinant of mRNA stability. *Cell*, 160(6), 1111–1124.

Prokopowich, C. D., Gregory, T. R., & Crease, T. J. (2003). The correlation between rDNA copy number and genome size in eukaryotes. *Genome*, 46(1), 48–50.

Puigbò, P., Bravo, I. G., & Garcia-Vallve, S. (2008). CAIcal: a combined set of tools to assess codon usage adaptation. *Biology direct*, 3(1), 38.

Qian, W., Liao, B.-Y., Chang, A. Y.-F., & Zhang, J. (2010). Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends in genetics*, 26(10), 425–430.

Qian, W. & Zhang, J. (2008). Gene dosage and gene duplicability. *Genetics*, 179(4), 2319–2324.

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., & Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS biology*, 4(10), e309.

Raper, K. B. (1935). Dictyostelium discoideum, a new species of slime mold from decaying forest leaves. *Journal of Agricultural Research*, 50(2), 135–147.

Raper, K. B. (1937). Growth and development of Dictyostelium discoideum with different bacterial associates. *Journal of Agricultural Research*, 55(4), 289–316.

Raper, K. B. (1940). Pseudoplasmodium formation and organization in Dictyostelium discoideum. *Journal of the Elisha Mitchell Scientific Society*.

Raper, K. B. (1951). Isolation, cultivation, and conservation of simple slime molds. *The Quarterly review of biology*, 26(2), 169–190.

Raser, J. M. & O'Shea, E. K. (2004). Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678), 1811–1814.

Ravarani, C. N. J., Chalancon, G., Breker, M., de Groot, N. S., & Babu, M. M. (2016). Affinity and competition for TBP are molecular determinants of gene expression noise. *Nature communications*, 7, 10417.

Reams, A. B. & Roth, J. R. (2015). Mechanisms of gene duplication and amplification. *Cold Spring Harbor perspectives in biology*, 7(2), a016592.

Rivero, F. (2002). mRNA processing in Dictyostelium: sequence requirements for termination and splicing. *Protist*, 153(2), 169–176.

Roberts, T. & Stewart, M. (1997). Nematode sperm: amoeboid movement without actin. *Trends in cell biology*, 7(9), 368–373.

Romaniuk, P. J., Lowary, P., Wu, H. N., Stormo, G., & Uhlenbeck, O. C. (1987). RNA binding site of R17 coat protein. *Biochemistry*, 26(6), 1563–1568.

Romans, P. & Firtel, R. A. (1985a). Organization of the actin multigene family of Dictyostelium discoideum and analysis of variability in the protein coding regions. *Journal of molecular biology*, 186(2), 321–335.

Romans, P. & Firtel, R. A. (1985b). Organization of the Dictyostelium discoideum actin multigene family. Flanking sequences show subfamily homologies and unusual dyad

symmetries. *Journal of molecular biology*, 183(3), 311–326.

Romans, P., Firtel, R. A., & Saxe, C. L. (1985). Gene-specific expression of the actin multigene family of Dictyostelium discoideum. *Journal of molecular biology*, 186(2), 337–355.

Romeralo, M., Cavender, J. C., Landolt, J. C., Stephenson, S. L., & Baldauf, S. L. (2011). An expanded phylogeny of social amoebas (Dictyostelia) shows increasing diversity and new morphological patterns. *BMC evolutionary biology*, 11, 84.

Rooney, A. P., Piontkivska, H., & Nei, M. (2002). Molecular evolution of the nontandemly repeated genes of the histone 3 multigene family. *Molecular biology and evolution*, 19(1), 68–75.

Rooney, A. P. & Ward, T. J. (2005). Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. *Proceedings of the National Academy of Sciences of the United States of America*, 102(14), 5084–5089.

Rosengarten, R. D., Santhanam, B., Fuller, D., Katoh-Kurasawa, M., Loomis, W. F., Zupan, B., & Shaulsky, G. (2015). Leaps and lulls in the developmental transcriptome of Dictyostelium discoideum. *BMC genomics*, 16, 294.

Rosengarten, R. D., Santhanam, B., & Kokosar, J. (2017). The Long Noncoding RNA Transcriptome of Dictyostelium discoideum Development. *G3: Genes*.

Rot, G., Parikh, A., Curk, T., Kuspa, A., Shaulsky, G., & Zupan, B. (2009). dictyExpress: a Dictyostelium discoideum gene expression database with an explorative data analysis web-based interface. *BMC bioinformatics*, 10, 265.

Rubin, R. W., Warren, R. H., Lukeman, D. S., & Clements, E. (1978). Actin content and organization in normal and transformed cells in culture. *The Journal of cell biology*, 78(1), 28–35.

Salje, J., Gayathri, P., & Löwe, J. (2010). The ParMRC system: molecular mechanisms of plasmid segregation by actin-like filaments. *Nature reviews. Microbiology*, 8(10), 683–692.

Sanchez, A., Choubey, S., & Kondev, J. (2013). Regulation of noise in gene expression. *Annual review of biophysics*, 42, 469–491.

Saran, S., Meima, M. E., Alvarez-Curto, E., Weening, K. E., Rozen, D. E., & Schaap, P. (2002). cAMP signaling in Dictyostelium. Complexity of cAMP synthesis, degradation and detection. *Journal of muscle research and cell motility*, 23(7-8), 793–802.

Sawai, S., Guan, X.-J., Kuspa, A., & Cox, E. C. (2007). High-throughput analysis of spatio-temporal dynamics in Dictyostelium. *Genome biology*, 8(7), R144.

Saxe, C. L., Johnson, R., Devreotes, P. N., & Kimmel, A. R. (1991). Multiple genes for cell surface cAMP receptors in Dictyostelium discoideum. *genesis*, 12(1-2), 6–13.

Sbalzarini, I. F. & Koumoutsakos, P. (2005). Feature point tracking and trajectory analysis

for video imaging in cell biology. *Journal of structural biology*, 151(2), 182–195.

Schafer, D. A. & Schroer, T. A. (1999). Actin-related proteins. *Annual review of cell and developmental biology*, 15(1), 341–363.

Schildmeyer, L. A., Braun, R., Taffet, G., Debiasi, M., Burns, A. E., Bradley, A., & Schwartz, R. J. (2000). Impaired vascular contractility and blood pressure homeostasis in the smooth muscle alpha-actin null mouse. *The FASEB Journal*, 14(14), 2213–2220.

Schmidt, E. E. & Schibler, U. (1995). Cell size regulation, a mechanism that controls cellular RNA accumulation: consequences on regulation of the ubiquitous transcription factors Oct1 and NF-Y and the liver-enriched transcription factor DBP. *The Journal of cell biology*, 128(4), 467–483.

Schmitt-Ney, M. & Habener, J. F. (2004). Cell-density-dependent regulation of actin gene expression due to changes in actin treadmilling. *Experimental Cell Research*, 295(1), 236–244.

Schnitzler, G. R., Briscoe, C., Brown, J. M., & Firtel, R. A. (1995). Serpentine cAMP receptors may act through a G protein-independent pathway to induce postaggregative development in Dictyostelium. *Cell*, 81(5), 737–745.

Schnitzler, G. R., Fischer, W. H., & Firtel, R. A. (1994). Cloning and characterization of the G-box binding factor, an essential component of the developmental switch between early and late development in Dictyostelium. *Genes & development*, 8(4), 502–514.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., & Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347), 337–342.

Segal, E. & Widom, J. (2009). Poly(dA:dT) tracts: major determinants of nucleosome organization. *Current opinion in structural biology*, 19(1), 65–71.

Shaner, N. C., Lambert, G. G., Chammas, A., Ni, Y., Cranfill, P. J., Baird, M. A., Sell, B. R., Allen, J. R., Day, R. N., Israelsson, M., Davidson, M. W., & Wang, J. (2013). A bright monomeric green fluorescent protein derived from Branchiostoma lanceolatum. *Nature methods*, 10(5), 407–409.

Sharon, D., Glusman, G., Pilpel, Y., Khen, M., Gruetzner, F., Haaf, T., & Lancet, D. (1999). Primate evolution of an olfactory receptor cluster: diversification by gene conversion and recent emergence of pseudogenes. *Genomics*, 61(1), 24–36.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., & Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature biotechnology*, 30(6), 521–530.

Sharon, E., van Dijk, D., Kalma, Y., Keren, L., Manor, O., Yakhini, Z., & Segal, E. (2014). Probing the effect of promoters on noise in gene expression using thousands of designed

sequences. *Genome research*.

Sharp, P. M. & Devine, K. M. (1989). Codon usage and gene expression level in Dictyostelium discoideum: highly expressed genes do 'prefer' optimal codons. *Nucleic acids research*, 17(13), 5029–5039.

Sharp, P. M. & Li, W. H. (1987). The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research*, 15(3), 1281–1295.

Shi, Y. & Manley, J. L. (2015). The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site. *Genes & development*, 29(9), 889–897.

Singh, G. P. (2013). Coupling between noise and plasticity in E. coli. *G3 (Bethesda, Md.)*, 3(12), 2115–2120.

Skerker, J. M., Perchuk, B. S., Siryaporn, A., Lubin, E. A., Ashenberg, O., Goulian, M., & Laub, M. T. (2008). Rewiring the specificity of two-component signal transduction systems. *Cell*, 133(6), 1043–1054.

Skoge, M., Wong, E., Hamza, B., Bae, A., Martel, J., Kataria, R., Keizer-Gunnink, I., Kortholt, A., van Haastert, P. J. M., Charras, G., Janetopoulos, C., & Irimia, D. (2016). A Worldwide Competition to Compare the Speed and Chemotactic Accuracy of Neutrophil-Like Cells. *PloS one*, 11(6), e0154491.

Skupsky, R., Burnett, J. C., Foley, J. E., Schaffer, D. V., & Arkin, A. P. (2010). HIV promoter integration site primarily modulates transcriptional burst size rather than frequency. *PLoS computational biology*, 6(9), e1000952.

Šlajcherová, K., Fiserova, J., Fischer, L., & Schwarzerová, K. (2012). Multiple actin isotypes in plants: diverse genes for diverse roles? *Frontiers in plant science*, 3, 226.

Small, E. C., Xi, L., Wang, J.-P., Widom, J., & Licht, J. D. (2014). Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity. *Proceedings of the National Academy of Sciences*, 111(24), E2462–71.

Sørensen, M. A., Kurland, C. G., & Pedersen, S. (1989). Codon usage determines translation rate in Escherichia coli. *Journal of molecular biology*, 207(2), 365–377.

Stark, R., Liebes, L. F., Nevrla, D., Conklyn, M., & Silber, R. (1982). Decreased actin content of lymphocytes from patients with chronic lymphocytic leukemia. *Blood*, 59(3), 536–541.

Stevense, M., Muramoto, T., Müller, I., & Chubb, J. R. (2010). Digital nature of the immediate-early transcriptional response. *Development (Cambridge, England)*, 137(4), 579–584.

Struhl, K. (1985). Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 82(24), 8419–8423.

Su, C., Jakobsen, I., Gu, X., & Nei, M. (1999). Diversity and evolution of T-cell receptor variable region genes in mammals and birds. *Immunogenetics*, 50(5-6), 301–308.

Suess, P. M. & Gomer, R. H. (2016). Extracellular Polyphosphate Inhibits Proliferation in an Autocrine Negative Feedback Loop in Dictyostelium discoideum. *Journal of Biological Chemistry*, 291(38), 20260–20269.

Suess, P. M., Watson, J., Chen, W., & Gomer, R. H. (2017). Extracellular polyphosphate signals through Ras and Akt to prime Dictyostelium discoideum cells for development. *Journal of cell science*, 130(14), 2394–2404.

Suganuma, T. & Workman, J. L. (2011). Signals and combinatorial functions of histone modifications. *Annual review of biochemistry*, 80, 473–499.

Sugino, R. P. & Innan, H. (2006). Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. *Trends in genetics*, 22(12), 642–644.

Sussman, R. & Sussman, M. (1967). Cultivation of Dictyostelium discoideum in axenic medium. *Biochemical and biophysical research communications*, 29(1), 53–55.

Suter, D. M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., & Naef, F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028), 472–474.

Svoboda, P. & Cara, A. D. (2006). Hairpin RNA: a secondary structure of primary importance. *Cellular and molecular life sciences*, 63(7-8), 901–908.

Swanson, J. A. & Taylor, D. L. (1982). Local and spatially coordinated movements in Dictyostelium discoideum amoebae during chemotaxis. *Cell*, 28(2), 225–232.

Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J., & Stahl, F. W. (1983). The double-strand-break repair model for recombination. *Cell*, 33(1), 25–35.

Szostak, J. W. & Wu, R. (1980). Unequal crossing over in the ribosomal DNA of Saccharomyces cerevisiae. *Nature*, 284(5755), 426–430.

Thattai, M. & van Oudenaarden, A. (2001). Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(15), 8614–8619.

Thomas, M. C. & Chiang, C.-M. (2006). The general transcription machinery and general cofactors. *Critical reviews in biochemistry and molecular biology*, 41(3), 105–178.

Tirosh, I. & Barkai, N. (2008). Two strategies for gene regulation by promoter nucleosomes. *Genome research*, 18(7), 1084–1091.

To, T.-L. & Maheshri, N. (2010). Noise can induce bimodality in positive transcriptional feedback loops without bistability. *Science*, 327(5969), 1142–1145.

Tomchik, K. J. & Devreotes, P. N. (1981). Adenosine 3',5'-monophosphate waves in Dictyostelium discoideum: a demonstration by isotope dilution–fluorography. *Science*, 212(4493), 443–446.

Trcek, T., Larson, D. R., Moldón, A., Query, C. C., & Singer, R. H. (2011). Single-molecule mRNA decay measurements reveal promoter- regulated mRNA stability in yeast. *Cell*, 147(7), 1484–1497.

Treisman, J., Gonczy, P., Vashishtha, M., Harris, E., & Desplan, C. (1989). A single amino acid can determine the DNA binding specificity of homeodomain proteins. *Cell*, 59(3), 553–562.

Tsang, A. S., Mahbubani, H., & Williams, J. G. (1982). Cell-type-specific actin mRNA populations in Dictyostelium discoideum. *Cell*, 31(2 Pt 1), 375–382.

Tsubota, S. I., Rosenberg, D., Szostak, H., Rubin, D., & Schedl, P. (1989). The cloning of the Bar region and the B breakpoint in Drosophila melanogaster: evidence for a transposon-induced rearrangement. *Genetics*, 122(4), 881–890.

Tuchman, J., Alton, T., & Lodish, H. F. (1974). Preferential synthesis of actin during early development of the slime mold Dictyostelium discoideum. *Developmental biology*, 40(1), 116–128.

Tvrdik, P. & Capecchi, M. R. (2006). Reversal of Hox1 gene subfunctionalization in the mouse. *Developmental cell*, 11(2), 239–250.

Uyemura, D. G., Brown, S. S., & Spudich, J. A. (1978). Biochemical and structural characterization of actin from Dictyostelium discoideum. *The Journal of biological chemistry*, 253(24), 9088–9096.

Valegård, K., Murray, J. B., Stockley, P. G., Stonehouse, N. J., & Liljas, L. (1994). Crystal structure of an RNA bacteriophage coat protein-operator complex. *Nature*, 371(6498), 623–626.

Van de Peer, Y., Mizrachi, E., & Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature reviews genetics*, 18(7), 411–424.

Van Driessche, N., Shaw, C., Katoh, M., Morio, T., Sucgang, R., Ibarra, M., Kuwayama, H., Saito, T., Urushihara, H., Maeda, M., Takeuchi, I., Ochiai, H., Eaton, W., Tollett, J., Halter, J., Kuspa, A., Tanaka, Y., & Shaulsky, G. (2002). A transcriptional profile of multicellular development in Dictyostelium discoideum. *Development (Cambridge, England)*, 129(7), 1543–1552.

van Haastert, P. J. M. (2010). Chemotaxis: insights from the extending pseudopod. *Journal of cell science*, 123(Pt 18), 3031–3037.

Vandekerckhove, J. & Weber, K. (1980). Vegetative Dictyostelium cells containing 17 actin genes express a single major actin. *Nature*, 284(5755), 475–477.

Varani, J., Wass, J. A., & Rao, K. M. (1983). Actin changes in normal human and rat leukocytes and in transformed human leukocytic cells. *Journal of the National Cancer Institute*, 70(5), 805–809.

Varnum, B., Edwards, K. B., & Soll, D. R. (1986). The developmental regulation of single-

cell motility in Dictyostelium discoideum. *Developmental biology*, 113(1), 218–227.

Vieira, F. G., Sánchez-Gracia, A., & Rozas, J. (2007). Comparative genomic analysis of the odorant-binding protein family in 12 Drosophila genomes: purifying selection and birth-and-death evolution. *Genome biology*, 8(11), R235.

Vorobiev, S., Strokopytov, B., Drubin, D. G., Frieden, C., Ono, S., Condeelis, J., Rubenstein, P. A., & Almo, S. C. (2003). The structure of nonvertebrate actin: implications for the ATP hydrolytic mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 100(10), 5760–5765.

Wachter, A. (2014). Gene regulation by structured mRNA elements. *Trends in genetics*, 30(5), 172–181.

Waddell, D. R. (1988). Cell size in Dictyostelium. *Developmental genetics*, 9(4-5), 673–681.

Walsh, J. B. & Stephan, W. (2008). Multigene Families: Evolution. John Wiley & Sons, Ltd, Chichester, UK.

Wapinski, I., Pfeffer, A., Friedman, N., & Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449(7158), 54–61.

Warner, J. R. (1999). The economics of ribosome biosynthesis in yeast. *Trends in biochemical sciences*, 24(11), 437–440.

Warrick, H. M. & Spudich, J. A. (1988). Codon preference in Dictyostelium discoideum. *Nucleic acids research*, 16(14A), 6617–6635.

Weake, V. M. & Workman, J. L. (2010). Inducible gene expression: diverse regulatory mechanisms. *Nature reviews genetics*, 11(6), 426–437.

Weiss, E. H., Mellor, A., Golden, L., Fahrner, K., Simpson, E., Hurst, J., & Flavell, R. A. (1983). The structure of a mutant H–2 gene suggests that the generation of polymorphism in H–2 genes may occur by gene conversion-like events. *Nature*, 301(5902), 671–674.

Wellauer, P. K. & Dawid, I. B. (1977). The structural organization of ribosomal DNA in Drosophila melanogaster. *Cell*, 10(2), 193–212.

Westphal, M., Jungbluth, A., Heidecker, M., Mühlbauer, B., Heizer, C., Schwartz, J. M., Marriott, G., & Gerisch, G. (1997). Microfilament dynamics during cell movement and chemotaxis monitored using a GFP-actin fusion protein. *Current biology*, 7(3), 176–183.

White, R. J. & Sharrocks, A. D. (2010). Coordinated control of the gene expression machinery. *Trends in genetics*, 26(5), 214–220.

Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., & Backofen, R. (2007). Inferring Noncoding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering. *PLoS computational biology*, 3(4), e65.

Williams, J. G. (2010). Dictyostelium Finds New Roles to Model. *Genetics*, 185(3), 717–726.

Wolf, L., Silander, O. K., & van Nimwegen, E. (2015). Expression noise facilitates the evolution of gene regulation. *eLife*, 4, 987.

Wu, M., Comeron, J. M., Yoon, H. S., & Bhattacharya, D. (2009). Unexpected dynamic gene family evolution in algal actins. *Molecular biology and evolution*, 26(2), 249–253.

Wu, X. & Bartel, D. P. (2017). Widespread Influence of 3'-End Structures on Mammalian mRNA Processing and Stability. *Cell*, 169(5), 905–917.e11.

Yant, L. & Bomblies, K. (2015). Genome management and mismanagement – cell-level opportunities and challenges of whole-genome duplication. *Genes & development*, 29(23), 2405–2419.

Yao, C., Biesinger, J., Wan, J., Weng, L., Xing, Y., Xie, X., & Shi, Y. (2012). Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. *Proceedings of the National Academy of Sciences*, 109(46), 18773–18778.

Yean, D. & Gralla, J. (1997). Transcription reinitiation rate: a special role for the TATA box. *Molecular and cellular biology*, 17(7), 3809–3816.

Yu, C.-H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M. S., & Liu, Y. (2015). Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Molecular cell*, 59(5), 744–754.

Yunger, S., Rosenfeld, L., Garini, Y., & Shav-Tal, Y. (2010). Single-allele analysis of transcription kinetics in living mammalian cells. *Nature methods*, 7(8), 631–633.

Zaki, M., Andrew, N., & Insall, R. H. (2006). Entamoeba histolytica cell movement: a central role for self-generated chemokines and chemorepellents. *Proceedings of the National Academy of Sciences of the United States of America*, 103(49), 18751–18756.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6), 292–298.

Zhou, M., Guo, J., Cha, J., Chae, M., Chen, S., Barral, J. M., Sachs, M. S., & Liu, Y. (2013). Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature*, 495(7439), 111–115.

Zhou, M., Wang, T., Fu, J., Xiao, G., & Liu, Y. (2015). Nonoptimal codon usage influences protein structure in intrinsically disordered regions. *Molecular microbiology*, 97(5), 974–987.

Zhou, T., Weems, M., & Wilke, C. O. (2009). Translationally optimal codons associate with structurally sensitive sites in proteins. *Molecular biology and evolution*, 26(7), 1571–1580.

Zhu, L., Zhang, Y., Hu, Y., Wen, T., & Wang, Q. (2013). Dynamic actin gene family evolution in primates. *BioMed research international*, 2013, 630803.

Zhurinsky, J., Leonhard, K., Watt, S., Marguerat, S., Bähler, J., & Nurse, P. (2010). A

coordinated global control over cellular transcription. *Current biology*, 20(22), 2010–2015.